BEYOND I, ROBOT: ETHICS, ARTIFICIAL INTELLIGENCE, AND THE DIGITAL AGE

VIRTUAL HEARING

BEFORE THE

TASK FORCE ON ARTIFICIAL INTELLIGENCE

OF THE

COMMITTEE ON FINANCIAL SERVICES

U.S. HOUSE OF REPRESENTATIVES

ONE HUNDRED SEVENTEENTH CONGRESS

FIRST SESSION

OCTOBER 13, 2021

Printed for the use of the Committee on Financial Services

Serial No. 117–52

U.S. GOVERNMENT PUBLISHING OFFICE
46–195 PDF WASHINGTON : 2022

HOUSE COMMITTEE ON FINANCIAL SERVICES

MAXINE WATERS, California, *Chairwoman*

CAROLYN B. MALONEY, New York
NYDIA M. VELÁZQUEZ, New York
BRAD SHERMAN, California
GREGORY W. MEEKS, New York
DAVID SCOTT, Georgia
AL GREEN, Texas
EMANUEL CLEAVER, Missouri
ED PERLMUTTER, Colorado
JIM A. HIMES, Connecticut
BILL FOSTER, Illinois
JOYCE BEATTY, Ohio
JUAN VARGAS, California
JOSH GOTTHEIMER, New Jersey
VICENTE GONZALEZ, Texas
AL LAWSON, Florida
MICHAEL SAN NICOLAS, Guam
CINDY AXNE, Iowa
SEAN CASTEN, Illinois
AYANNA PRESSLEY, Massachusetts
RITCHIE TORRES, New York
STEPHEN F. LYNCH, Massachusetts
ALMA ADAMS, North Carolina
RASHIDA TLAIB, Michigan
MADELEINE DEAN, Pennsylvania
ALEXANDRIA OCASIO-CORTEZ, New York
JESÚS "CHUY" GARCIA, Illinois
SYLVIA GARCIA, Texas
NIKEMA WILLIAMS, Georgia
JAKE AUCHINCLOSS, Massachusetts

PATRICK McHENRY, North Carolina,
Ranking Member
FRANK D. LUCAS, Oklahoma
BILL POSEY, Florida
BLAINE LUETKEMEYER, Missouri
BILL HUIZENGA, Michigan
ANN WAGNER, Missouri
ANDY BARR, Kentucky
ROGER WILLIAMS, Texas
FRENCH HILL, Arkansas
TOM EMMER, Minnesota
LEE M. ZELDIN, New York
BARRY LOUDERMILK, Georgia
ALEXANDER X. MOONEY, West Virginia
WARREN DAVIDSON, Ohio
TED BUDD, North Carolina
DAVID KUSTOFF, Tennessee
TREY HOLLINGSWORTH, Indiana
ANTHONY GONZALEZ, Ohio
JOHN ROSE, Tennessee
BRYAN STEIL, Wisconsin
LANCE GOODEN, Texas
WILLIAM TIMMONS, South Carolina
VAN TAYLOR, Texas
PETE SESSIONS, Texas

Charla Ouertatani, *Staff Director*

TASK FORCE ON ARTIFICIAL INTELLIGENCE

BILL FOSTER, Illinois, *Chairman*

BRAD SHERMAN, California
SEAN CASTEN, Illinois
AYANNA PRESSLEY, Massachusetts
ALMA ADAMS, North Carolina
SYLVIA GARCIA, Texas
JAKE AUCHINCLOSS, Massachusetts

ANTHONY GONZALEZ, Ohio, *Ranking Member*
BARRY LOUDERMILK, Georgia
TED BUDD, North Carolina
TREY HOLLINGSWORTH, Indiana
VAN TAYLOR, Texas

CONTENTS

Page

Hearing held on:
October 13, 2021 ... 1
Appendix:
October 13, 2021 ... 25

WITNESSES

Wednesday, October 13, 2021

Broussard, Meredith, Associate Professor, Arthur L. Carter Journalism Institute of New York University ... 4
Cooper, Aaron, Vice President, Global Policy, BSA—The Software Alliance ... 12
King, Meg, Director, Science and Technology Innovation Program, The Wilson Center ... 8
Vogel, Miriam, President and CEO, EqualAI ... 6
Yong, Jeffery, Principal Advisor, Financial Stability Institute, Bank for International Settlements ... 10

APPENDIX

Prepared statements:
Broussard, Meredith ... 26
Cooper, Aaron ... 33
King, Meg ... 74
Vogel, Miriam ... 79
Yong, Jeffery ... 88

Additional Material Submitted for the Record

Garcia, Hon. Sylvia:
Written responses to questions for the record submitted to Jeffery Yong ... 90

BEYOND I, ROBOT: ETHICS, ARTIFICIAL INTELLIGENCE, AND THE DIGITAL AGE

Wednesday, October 13, 2021

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 12 p.m., via Webex, Hon. Bill Foster [chairman of the task force] presiding.

Members present: Representatives Foster, Casten, Pressley, Adams, Garcia of Texas, Auchincloss; Gonzalez of Ohio, Loudermilk, Budd, and Taylor.

Chairman Foster. The Task Force on Artificial Intelligence will 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.

As a reminder, I ask all Members to keep themselves muted when they are not being recognized by the Chair. The staff has been instructed not to mute Members, except when a Member is not being recognized by the Chair and there is inadvertent background noise.

Members are also reminded that they may only participate in one remote proceeding at a time. If you are participating today, please keep your camera on, and if you choose to attend a different remote proceeding, please turn your camera off.

Today's hearing is entitled, "Beyond I, Robot: Ethics, Artificial Intelligence, and the Digital Age."

I now recognize myself for 4 minutes to give an opening statement.

Thank you, everyone, for joining us today at a time when the power and perils of artificial intelligence (AI) are very much on people's minds. Each generation has its own cautionary tales about AI. Recent big-screen adaptations—The Matrix, Terminator, and Tron—echo the episodes of the old 1960s Star Trek starring William Shatner as Captain James D. Kirk of the Starship Enterprise, and those episodes themselves were taken from the short stories of Isaac Asimov, Arthur Clarke, and all of the old masters of 1950s sci-fi pulp magazines. And parenthetically, I should offer our congratulations that today, William Shatner was able to boldly go into suborbital near space where X–15 pilots have been boldly going since the late 1950s. But I digress. Asimov's classic, I, Robot, showed us what can happen when we deploy technology or AI without fully comprehending its consequences.

There is an ancient joke in AI that I first heard as an undergraduate back in the 1970s about an all-powerful AI that was given a simple command: Maximize paperclip production. It thought about it for a moment and then began killing off all humans on Earth because humans interfere with paperclip production.

Now, it may have taken us 50 years, but we are kind of there. Facebook's AI was given the simple command, "maximize Facebook's profits," whereupon they thought for a moment and then began killing off all rational political debate in our country because that interferes with Facebook's profits. And the situations with social media in Myanmar and around the world are even uglier and more deadly.

In previous hearings of this task force, we have looked at the biases and unexpected side effects of using AI in financial services and housing. We have also looked at the implications of artificial intelligence's voracious appetite for personal data and the implications for privacy, at technological approaches to maximally preserve privacy while retaining AI's effectiveness, and the importance of secure digital identity.

In this hearing, we are going to take a closer look at the frameworks for developing, monitoring, and recognizing AI to ensure that the technology we develop and deploy will be of overall benefit to society.

In past hearings, we have examined instances of algorithmic bias that have produced discriminatory effects in the lending space. We have seen facial recognition technology that is far less effective at identifying minorities correctly, despite the fact that the developers of these tools did not include a discriminatory line of code in their products. So, we clearly cannot allow technology to treat humans differently based on race and appearance, unless, of course, perhaps we are explicitly correcting for past unjust biases, which brings up a set of issues that my father struggled with as a civil rights lawyer back in the 1950s, and continues with us today. We have to understand whether we should hold AI to standards that are higher than we would expect of an ordinary human-based decision-making process.

As we start defining frameworks for developing and performance-testing AI, it seems possible that we are starting to place requirements on AI that are more strict than we would ever place on human decision-makers. For example, most of our witnesses today have advocated for defining minimum diversity standards for the training datasets for AI, but we have never considered requiring that a human bank officer would have a minimum number of friends of different races or protected classes, even though it might arguably result in more fair decision-making. And we may already be seeing the positive results of holding AI to higher standards than humans with the recent reports that fintech apps were apparently more effective than human-based banks in issuing Paycheck Protection Program (PPP) loans to minority customers.

As policymakers, we also have to understand to what extent we should concentrate on so-called black-box testing that only focuses on the inputs and outputs from opaque neural networks and other decision-making algorithms, or whether we should expect ourselves and the public to receive and to understand a detailed explanation of what goes on under the hood. So, there is a lot to be examined here. It is my hope that in this dialogue, we will discover which frameworks exist and which should be created or fleshed out to ensure that AI is working effectively and safely for everyone.

And the Chair will now recognize the ranking member of the task force, Mr. Gonzalez of Ohio, for 5 minutes for an opening statement.

Mr. GONZALEZ OF OHIO. Thank you, Chairman Foster, for your leadership on ethics in AI and for convening today's hearing, and I also thank our witnesses for being here. It is vital that Congress continues to consider how we can best promote innovative advancement in the private sector while also ensuring that AI is both transparent and ethical. Today's hearing provides an opportunity to hear directly from industry experts and stakeholders on the importance of this topic.

A few months ago, the task force held a similar hearing examining how human-centered AI can address systemic racism. One of our witnesses at that hearing, Professor Rayid Ghani of Carnegie Mellon University, testified that algorithms themselves are neither inherently biased or unbiased, but work by analyzing past data and making generalizations about future outcomes. I believe that these discussions on bias and algorithms are important to have. We must acknowledge and recognize these technologies at times are not perfect due to the inherent nature of a technology created by humans. It is vital, though, that we do not take steps backwards by overregulating this industry, which may have a chilling effect on the deployment of these technologies.

If there are problems with AI and algorithms, we should not abandon our push to innovate and move forward. It is through further innovation that we are likely going to be able to fix these issues and to improve the technology. As Chairman Foster recognized, we have seen the benefits in the disbursement of PPP loans. I think that is an important thing for us to keep in mind as we continue forward.

We should also continue to work with the experts in industry in order to move forward in a bipartisan way that both celebrates technical advancements and ensures that there is transparency and fairness through the use of artificial intelligence. There have been multiple efforts in the government and the private sector to address this issue, and we have seen tremendous advances not only in AI technology, but in efforts to address bias in algorithms internally. There is recognition of a business incentive to have transparent algorithms that are fair and ethical.

Beyond the obvious concerns of ethics and transparency, I am also looking forward to learning more today from our witnesses about ways that we can strengthen data transparency for families, and consider reforms that would protect our children from being targeted by harmful algorithms. As the financial internet and the traditional internet merge—and we have seen recently-reported social media companies, like TikTok, employing algorithms that promote inappropriate content to young users—I think it is extremely troubling and extremely timely that we start to discuss these things. An AI-powered world where parents have no control over what content or products are being fed to their kids, no transparency around the algorithms that are funneling the content, and no control over the underlying data itself is not an ideal outcome.

In summary, AI has great promise to innovate industries like the financial services sector, but there are still opportunities to improve. I look forward to hearing from our witnesses today how Congress should be thinking about this balance, and I yield back.

Chairman FOSTER. Thank you.

The Chair will now recognize the Chair of the full Financial Services Committee, the gentlewoman from California, Chairwoman Waters, for 1 minute.

[No response.]

Chairman FOSTER. It is my understanding that she is not able to make it right now, so we will move on.

Today, we welcome the testimony of our distinguished witnesses: Ms. Meredith Broussard, an associate professor at the Arthur L. Carter Journalism Institute of New York University; Ms. Miriam Vogel, the president and CEO of EqualAI; Ms. Meg King, the director of the Science and Technology Innovation Program at the Wilson Center; Mr. Jeffery Yong, principal advisor at the Financial Stability Institute of the Bank for International Settlements; and Mr. Aaron Cooper, the vice president for global policy at BSA—The Software Alliance.

Witnesses are reminded that their oral testimony will be limited to 5 minutes. You should be able to see a timer on your screen which indicates how much time you have left. I would ask you to be mindful of the timer, and quickly wrap up your testimony once the time has expired, so that we can be respectful of both the witnesses' and the members' time.

And without objection, your written statements will be made a part of the record.

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

STATEMENT OF MEREDITH BROUSSARD, ASSOCIATE PROFESSOR, ARTHUR L. CARTER JOURNALISM INSTITUTE OF NEW YORK UNIVERSITY

Ms. BROUSSARD. Thank you. Chairman Foster, members of the task force, thank you for hosting this important hearing and for giving me the opportunity to testify. My name is Meredith Broussard. I am a professor at NYU, the research director at the NYU Alliance for Public Interest Technology, and author of the book, "Artificial Unintelligence: How Computers Misunderstand the World." In my written testimony, I explore a practical vision for recognizing AI, and in my short time, I'll talk about AI generally as well as discrimination algorithmic auditing and regulatory sandboxes.

The first thing I want to say is that AI is not what we see in Hollywood. There is no robot apocalypse coming. There is no singularity. We do not need to prepare for artificial general intelligence because these things are imaginary. What is real is that AI is math, very complicated and beautiful math. Machine learning, the most popular kind of AI, is a poorly-chosen term because it suggests that there is a brain or sentience inside the computer. There is not. When we do machine learning, we take a large set of historical data and instruct the computer to create a model based on patterns and values in that dataset. The model can then be used to predict or make decisions based on past data. The more data you put in, the more precise your predictions will become. However, all historical datasets have bias. For example, if you feed in data on who has gotten a mortgage in the past in the United States and ask the computer to make similar decisions in the future, you will get an AI that offers mortgages to more White people than people of color.

AI needs to be regulated because it has all of the flaws of any human process, plus some. My own regulatory vision begins with frameworks, high-level governance models that guide a company's use of AI and data. A company can make sure its frameworks are implemented by performing regular algorithmic audits, ideally using a regulatory sandbox. The process could be monitored by regulators using tools we already have, namely compliance processes inside existing regulatory agencies. Agencies and companies might decide which AIs need to be regulated and monitored by looking at the user and the context. Automated license plate readers used at toll booths might be a reasonable use of AI. Automated license plate readers used by police as dragnet surveillance might be an unreasonable use of AI.

An open secret in the AI world is everyone knows that these systems discriminate. Any conversation about a robot apocalypse is a deliberate distraction from the harms that AI systems are causing today. Right now, AI is preventing people from getting mortgages. A recent investigation by The Markup found that nationally, loan applicants of color were 40 to 80 percent more likely to be turned down by mortgage approval algorithms as compared to their White counterparts.

When the International Baccalaureate used AI to assign student grades during the pandemic, high-achieving, low-income students received terrible grades, which prevented them from getting college credits that would allow them to graduate early and incur less student loan debt.

AI is used to generate secret predictive consumer scores, like health risk scores or identity and fraud scores. It is likely that Black, Indigenous, and People of Color (BIPOC) people are systematically disadvantaged by most of these scoring systems. The EU's proposed AI regulation calls for categorizing AI into high and low risk, which I think is a good strategy. A low-risk use might be using facial recognition to unlock your phone. A high-risk use might be the police using facial recognition on real-time surveillance video feeds. Facial recognition has been shown to consistently misidentify people with darker skin; people of color are at a high risk of being harmed by facial recognition when it is used in policing. In the U.S., we can register and audit high-risk AI to ensure that AI is not harming citizens.

The process for uncovering algorithmic bias is called algorithmic auditing. ORCAA, a company I consult with, performs bespoke algorithmic audits in context, asking how an algorithm might fail and for whom. Audits can show how an algorithm might be racist, or sexist, or ableist, or might discriminate illegally. Once we identify a problem, it can be addressed, or the algorithm can be discarded. There is also software like Parity, or Aequitas, or AI Fairness 360, that can evaluate algorithms for 1 of 21 known kinds of mathematical fairness.

I'm enthusiastic about the potential of a regulatory sandbox, a protected environment where companies can test their algorithms for bias. If and when the bias is discovered, they can then address the issue in their code and rerun the test until they're in compliance with acceptable thresholds. I'm currently working with ORCAA to develop a regulatory sandbox prototype. In our version, regulators would also have a limited view inside the sandbox to see if the company is auditing their algorithms for bias and fixing the problems that they find without the companies revealing any trade secrets.

Thank you for the opportunity to testify today on this important topic, and I welcome your questions.

[The prepared statement of Ms. Broussard can be found on page 26 of the appendix.]

Chairman FOSTER. Thank you, Ms. Broussard, and I have to say I am fascinated with the thought of figuring out for which of the 21 definitions of fairness you will be advocating.

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

STATEMENT OF MIRIAM VOGEL, PRESIDENT AND CEO, EQUALAI

Ms. VOGEL. Chairman Foster, Ranking Member Gonzalez, and distinguished members of the task force, thank you for conducting this important hearing and for the opportunity to provide this testimony. My name is Miriam Vogel. I'm president and CEO of EqualAI, a nonprofit founded to reduce unconscious bias in AI systems. At EqualAI, we are AI net positive. We believe AI is and will be a powerful tool to advance our lives, economy, and opportunities to thrive, but only if we're vigilant to ensure that the AI we use does not perpetuate and mass produce historical and new forms of bias and discrimination.

We're at a critical juncture. AI is increasingly becoming an important part of our daily lives, but decades of progress made and lives lost to promote equal opportunity can be unwritten in a few lines of code. And the perpetrators of this disparity may not even realize the harm they're causing. For instance, we can see our country's long history of housing discrimination now replicated at scale in mortgage approval algorithms that determine creditworthiness using proxies for race and class.

At EqualAI, we try to help avoid such harms by supporting three main stakeholders: companies; policymakers; and lawyers. Often, our work involves helping organizations understand they are effectively AI companies because they are now using AI in pivotal functions. As such, they need an AI governance plan, particularly given that with AI, as you know, key assessments occur behind the proverbial black box where inputs and operations are generally unknown to the end user.

As discussed in your past hearings, implicit bias infiltrates AI in a variety of ways. Our operating thesis is that bias can embed in each of the human touch points throughout the AI lifecycle, from the ideation phase deciding what the problem is you even want to solve with AI, to the design, data collection, development, testing, and monitoring phases. But we are optimistic and we think each touch point is also an opportunity to identify and eliminate harmful biases. As such, risk management should occur at each stage of the AI lifecycle.

There are several helpful frameworks to identify and reduce harms in the AI systems, including GAO's, GSA's, and the important efforts under way at NIST. The EqualAI framework offers five pillars to consider when establishing responsible governance, including, first, invest in the pipeline. Our basis tenet is that AI needs to be created by and for a broader cross-section of our population. There are several organizations promoting diversity in tech effectively right now—AINU, AI4All, and several others—and we need to support these efforts.

Second, hire and promote people with your values. To create and sustain a diverse workplace and produce better AI, AI programs used in H.R. functions should be checked routinely to ensure they're in sync with the values of your organization and our country.

Third, evaluate your data. The more we know about datasets, the safer we are as a society. We encourage identifying gaps in data so that they can be rectified and, at a minimum, clarified for end users.

Fourth, test your AI. AI should be checked for bias on a routine basis. As you know, AI constantly iterates and learns new patterns as it is fed new data. On our website, EqualAI.org, we offer a checklist to help get you started, and we offer additional steps to take in our written testimony. We highly recommend as well the use of routine audits.

And fifth, redefine the team. An often-overlooked opportunity to reduce bias in AI is by creating testing teams that include those underrepresented in the AI creation and the underlying datasets.

There are numerous ways that Congress can play a key role in ensuring more effective, inclusive AI. Several are listed in our testimony. A few include, one, Congress can reinforce the applicability of laws prohibiting discrimination to AI-supported determinations. Two, Congress can lead by example, create a framework for AI procurement, acquisition, and development and ask vendors if they do the same.

Three, incentivize investment in the future of work. Like all transformative technologies, AI will eliminate jobs, but it will also open up opportunities. To lead in the AI revolution, safeguard our economy, and support greater prosperity among more communities, we should re-skill our workforce by understanding what jobs are likely to emerge, and offering incentives for upscaling and loan forgiveness for those committing to a term in public service.

Finally, we enthusiastically support the bill of rights put forward by the White House Office of Science and Technology Policy last week to level-set expectations and inform the public about their rights.

In conclusion, we believe we're at a critical juncture to ensure that AI is built by and for a broader cross-section of our population. It's not only the right thing to do; a strong U.S. economy and our leadership depend on it. Thank you for the opportunity to testify, and I look forward to your questions.

[The prepared statement of Ms. Vogel can be found on page 79 of the appendix.]

Chairman FOSTER. Thank you, Ms. Vogel, and I echo your enthusiasm for the White House's effort to come up with an AI bill of rights, though I don't believe I have seen even a draft of it at this point.

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

STATEMENT OF MEG KING, DIRECTOR, SCIENCE AND TECHNOLOGY INNOVATION PROGRAM, THE WILSON CENTER

Ms. KING. Thank you, Chairman Foster, Ranking Member Gonzalez, and members of the AI Task Force for inviting me to testify today. My name is Meg King. I'm the director of the Science and Technology Innovation Program at the Wilson Center, a nonpartisan think tank created by Congress nearly 60 years ago. My program studies the policy opportunities and challenges of emerging technologies, and investigates methods to foster more open science and to build serious games. We also offer hands-on training programs, called the Technology Labs, to Legislative and Executive Branch staff on a variety of issues, including artificial intelligence. Next month, we will offer a series of individual trainings on AI for Members as well.

As with any technological evolution, the benefits of AI come with associated costs and risks. Focusing only on the benefits misses the nuances of the potentials and pitfalls of this advance. To help the task force 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.

Today, there aren't 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 organizations, regions, and private companies have all begun to issue ethical frameworks for AI. Most are very vague principles, as you mentioned, Chairman Foster, with little guidance as to application.

Two that this committee should pay close attention to are the Organisation for Economic Co-operation and Development (OECD), and the European Commission (EC). In addition to their principles on AI, the OECD is developing process and technical guidelines ranging from pinpointing new research to making available software advances which will become part of a publicly-available interactive tool for developers and policymakers alike. As Ms. Broussard noted, European regulators announced a risk-based plan this year to establish transparency requirements, 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.

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 not addressed properly. 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, specifically for the financial sector, should include explainability, data inputs, testing, and system life cycle. Explainable Artificial Intelligence (XAI) is the method to ask questions about the outcomes of AI systems and how they achieve them. It helps developers and policymakers identify problems and failures, possible sources of bias, and helps users access explanations. There are a number of techniques available to carry out XAI, as well as open source tools, which make these techniques more accessible.

In the financial sector, XAI will become critical as predictive models increasingly perform calculations during live transactions, for example, to evaluate risk or the opportunity of offering a financial product or specific transaction to a customer. Establishing a clear process for XAI 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 on how these systems will be used and in what context will be helpful. Without context, data pulled from a mix of public/private records can produce inaccurate results and discriminate in access to financial products. One of the near-term questions this committee should ask about systems you will encounter in your oversight is how the COVID-19 pandemic experience is factored into these systems. One promising possibility to address the data input problem might be to synthesize artificial financial data to correct for inaccurate or biased historical data. Just today, a major tech company announced acquisition of a synthetic data startup. Watch this space.

While quality assurance is part of most development processes, there are currently no enforceable standards for testing AI systems, and, therefore, testing is uneven at best. Additionally, users are far removed from AI system developers. Carefully assessing the growing field of Machine Learning Operations Tools (MLOps) and machine learning operations and identifying ways the committee can participate in that process will be useful.

AI breaks, often in unpredictable ways, at unpredictable times. Participants in the Wilson Center's AI Lab have seen AI function spectacularly 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 length of purchase.

Beyond mistakes, some AI systems carry out tasks in a way humans never would. Many examples exist of scenarios producing results developers didn't intend, like a vacuum cleaner injecting collected dust so it can collect even more, and a racing boat in a digital game looping in place to collect points instead of winning the race. Anyone who has played the game, "20 Questions" understands this problem. Unless you ask exactly the right question, you won't get the right answer.

As more and more AI systems are built and distributed widely with varying levels of user expertise, this problem will 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. Thank you.

[The prepared statement of Ms. King can be found on page 74 of the appendix.]

Chairman FOSTER. Thank you, Ms. King.

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

STATEMENT OF JEFFERY YONG, PRINCIPAL ADVISOR, FINANCIAL STABILITY INSTITUTE, BANK FOR INTERNATIONAL SETTLEMENTS

Mr. YONG. Thank you. Good afternoon, Chairman Foster, Ranking Member Gonzalez, and distinguished members of the task force. My name is Jeffery Yong, and I'm the principal advisor at the Financial Stability Institute of the Bank for International Settlements, or the BIS. I offer my remarks today entirely in my personal capacity based on a publication that I co-authored with my colleague, Jermy Prenio, entitled, "FSI Insights No. 35: Humans keeping AI in check—emerging regulatory expectations in the financial sector." And the views expressed in that paper are our own and do not necessarily represent those of the BIS, its members, or the Basel ommittees. I'm appearing before the task force voluntarily. I would like to note that my statements here today are similarly my personal views, and they do not represent the official views of the BIS, its members, or the Basel Committees.

By way of background, the Financial Stability Institute (FSI) is a unit within the BIS with a mandate to support implementation of global regulatory standards and sound supervisory practices by central banks and financial sectors, supervisory and regulatory authorities worldwide. One of the ways the FSI carries out this mandate is through its policy implementation work which involves publishing FSI Insights papers. The papers aim to contribute to international discussions on a range of contemporary, regulatory, and supervisory policy issues and implementation challenges faced by financial sector authorities.

In preparing FSI Insight No. 35, my co-author and I found that regulatory expectations on the use of AI in financial services were at a nascent stage. Accordingly, we drafted a paper with four key objectives: to identify emerging common financial regulatory themes around AI governance; to assess how similar or different these common regulatory themes are viewed in the context of AI vis-a-vis that of traditional financial models; to explore how existing international financial regulatory standards may be applied in the context of AI governance; and to examine challenges in implementing the common regulatory themes.

To this end, we can select a section of policy documents on AI governance issued by financial authorities or groups formed by them as well as other cross-industry AI governance guidance that applies to the financial sector. In total, we examined 19 policy documents issued by 16 regional and national authorities and 2 international organizations. Most of these documents are either discussion papers or high-level principles, which underscores the fact that financial regulatory thinking in this area is at a very early stage.

We identified five common themes that recur in policy documents that we examined: reliability; accountability; transparency; fairness; and ethics.

On the theme of reliability, emerging supervisory expectations for AI and traditional models appear to be similar. What seems to be different is that the reliability of AI models is viewed from the perspective of avoiding harm to data subjects or consumers, for example, through discrimination.

On the theme of accountability, it is acknowledged that both traditional and AI models require human intervention. In the case of AI, however, this requirement is motivated by the need to make sure that decisions based on AI models do not result in unfair or unethical outcomes. Moreover, external accountability is emphasized in the case of an AI model, so that data subjects are aware of AI-driven decisions and have channels for recourse and moving on transparency.

Supervisory expectations related to explainability and auditability are similar for AI and traditional models. However, expectations or external disclosure are unique to AI models. This refers to expectations that firms using AI models should make data subjects aware of AI-driven decisions that impact them, including how their data is being used.

On the theme of fairness, there's a distinct and strong emphasis in emerging supervisory expectations on this aspect in the case of AI models. Fairness is commonly described in the documents as avoiding discriminatory outcomes.

Similarly, on ethics, as a distinct and strong emphasis on this aspect of AI models, ethics expectations are broader than fairness, and relate to ascertaining that consumers will not be exploited or harmed.

Now, given the similarities of the themes between AI and traditional models, existing financial literacy standards that govern the use of traditional models may be applied in the context of AI. However, there may be scope to do more in defining financial regulatory expectations related to fairness and ethics. The use of AI in the financial sector presents certain challenges, and the key challenge relates to the level of complexity and lack of explainability. Given these challenges, one way to approach this is to consider a tailored and coordinated regulatory policy approach, meaning differentiating potential and conduct treatment depending on the risk that the AI models pose.

With that, I conclude. Thank you.

[The prepared statement of Mr. Yong can be found on page 88 of the appendix.]

Chairman FOSTER. Thank you, Mr. Yong.

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

STATEMENT OF AARON COOPER, VICE PRESIDENT, GLOBAL POLICY, BSA—THE SOFTWARE ALLIANCE

Mr. COOPER. Thank you very much. Good afternoon, Chairman Foster, Ranking Member Gonzalez, and members of the AI Task Force. My name is Aaron Cooper. I'm vice president of global policy for BSA-The Software Alliance. BSA is the leading advocate for the global enterprise software industry. Our members are at the forefront of developing cutting-edge, data-driven services that have a significant impact on U.S. job creation. I commend the task force for convening today's important hearing, and I thank you for the opportunity to testify.

Enterprise software services, including AI, are accelerating digital transformation in every sector of the economy, and BSA members are on the leading edge, providing businesses with the trusted tools they need to leverage the benefits of AI. In fact, last year, software supported more than 12.5 million jobs outside the tech sector. AI is not just about robots, self-driving vehicles, or social media. It's used by businesses of all sizes to improve their competitiveness. It's the power and industrial design that improves manufacturing performance and reduces environmental impact. It's the tool that streamlines transportation and logistics operations, and that detects cyberattacks and improves H.R. operations. In the financial services industry, AI is being used to reduce the risk of fraudulent transactions and deliver a better customer relations experience.

While the adoption of AI can unquestionably be a force for good, it can also create real risks if not developed and deployed responsibly. We commend the task force for its work to explore domestic and international AI frameworks because they play a critical role in ensuring the responsible use of AI.

As you explore these issues, we offer our perspective on a risk management approach to bias which has been a particular focus for BSA, and that we hope will also inform the broader conversation. For BSA members, earning trust and confidence in AI and other software services they develop is crucial, so confronting the risk of bias is a priority. We, therefore, set out to develop concrete steps companies can take to guard against this. The resulting framework is included in full in my written testimony. It is built on three key elements: impact assessments; risk mitigation practices; and organizational accountability modeled on NIST frameworks, which includes more than 50 actionable diagnostic statements for performing impact assessments that identify risks of bias and corresponding best practices for mitigating those risks.

Among the unique features of the BSA framework is that it recognizes that these steps need to be followed at all stages of the AI life cycle: design; development; and deploymentt. Also, different businesses will have different roles throughout the life cycle, so risk management responsibilities will need to be tailored to a company's role. Who's developing the algorithm? Who's collecting the data, training the model, and ultimately deploying the system? What does that all mean in practice?

A few examples. First, when designing an AI system, companies should clearly define the intended use and what the system is optimized to predict, identify who may be impacted, and, if the risk of bias is present, document efforts to mitigate that risk. They should examine data that will be used to train the model to ensure that it's representative and not tainted by historical biases.

Second, at the development stage, they should document choices made in selecting features for the model and document how the model was tested.

Third, at the deployment phase, they should document the process for monitoring the data and model and maintain a feedback mechanism to enable consumers to report concerns.

And to be clear, at every phase, it is important for companies to have a team that brings diverse perspectives and background, which can help anticipate the needs and concerns of people who may be affected by AI in order to identify potential sources of bias. Bias is only one of the important ethical considerations for responsible AI, but addressing it is critical. And the risk management approach we recommend in this context can be tailored to address other ethical considerations.

In conclusion, digital transformation across industry sectors is creating jobs and improving our lives, but industry, civil society, and academia must work together with Congress and other policymakers on guidelines and laws which will ensure that companies act responsibly in how they develop and deploy AI. We appreciate the task force's strong focus on these issues and hope that our framework on confronting bias will contribute meaningfully to this discussion. Thank you for the opportunity to testify, and I look forward to your questions.

[The prepared statement of Mr. Cooper can be found on page 33 of the appendix.]

Chairman FOSTER. Thank you, Mr. Cooper.

I will now recognize myself for 5 minutes for questions.

My first general question is to Ms. King or Mr. Cooper, whomever wants to field it. How much should we expect of AI, and, in particular, should we be asking more of AI than we do of humans? For AI-driven cars, should the standard be that you should outperform humans on average or in all circumstances? With similar things regarding fairness as well, in general, is it reasonable? Are there real dangers in using human-based decision-making as the standard of fairness and safety for what is acceptable in AI?

Mr. COOPER. I am happy to jump in. I will give one example and a way of thinking about this, that things which are illegal in the physical world should be illegal in the digital world when we use AI or any other system. In the realm of discrimination, for instance, a practice that would be discriminatory if a person did it, should still be discriminatory and illegal if an AI system does it.

And I think what we are finding in other areas is that AI is increasingly being used, both in everyday features of what companies are doing as they go through a digital transformation but will also increasingly be used in more high-risk areas. And in those situations, we need to make sure that there is a proper impact assessment so we know, whether it is related to bias or safety or another issue, that companies are thinking through what those implications are going to be and taking steps to mitigate the risks.

Ms. VOGEL. I am happy to answer if you would like, as well. I think the answer is, honestly, we don't know. Certain systems are designed very narrowly right now, and that is because AI outperforms humans in those systems. But in others, with context, with heuristics, the shortcuts that we use as humans don't perform well. And as one of the Wilson Center machine-learning researchers who has written a paper about it reminds us regularly, autonomous agents optimize the room, floor, and function that we give them. So, until we can improve AI to a level where we feel comfortable that moves beyond that narrow capability, I don't think we have an answer yet about how to think about the consequences, but also the opportunities. They are just so varied across so many sectors at this point.

Chairman FOSTER. Are there any other comments on the deployment decision that has to be made here? You need some sort of absolute standard that this is good enough for this application, and it is something we are going to have to pace because that is probably, at best, the level at which Congress will be specific about how these decisions should be set up.

Another thing that I know we all struggle with is this question of black-box testing versus expecting that the public should have a detailed understanding of what goes on inside. If you look at the trouble that we have had trying to convince people to get vaccinated, it is not clear that it helps to tell them the details of how the immune system in the human works. And, that may make it better. It may make it worse. We had this situation very recently where we apparently fired a football coach, not, to my knowledge, for mistreating athletes, but for what went on in his private decision-making.

Should we accept or reject algorithms based only on their inputs and outputs, or should we actually demand to look inside at all of the intermediate levels of the neural network and see if there are objectionable racist nodes in them? What is your thinking on that, the black box versus detail, and also how to convey that to the public? Anyone? Should I just pick someone at random?

Ms. KING. I am happy to jump in.

Chairman FOSTER. Okay.

Ms. KING. I think it is a great question, and I think that the challenge also includes that even if you show the general public all of the nodes, it wouldn't necessarily make sense. In this case, you wouldn't know which are prioritized, so there is a balance to strike. There are intellectual property issues, privacy issues, and so forth. So, just opening the box, first of all, would be somewhat technically challenging as well as legally. Compliance testing, as you say, can be a helpful way to demonstrate compliance, safety, and legality. And to the extent that more data becomes available, we don't need to expect everyone in the general public to understand it. We have seen so many cases already where the limited publicly-available data has been used for important findings, like with the UnitedHealth Care Optum case, where scientists, researchers were able to go backwards, look at the algorithms, and identify biases in the algorithms.

Chairman FOSTER. Thank you. And when you figure all this stuff out, let us know.

The Chair will now recognize the ranking member of the task force, Representative Gonzalez of Ohio, for 5 minutes.

Mr. GONZALEZ OF OHIO. Thank you, Chairman Foster, and thank you to our witnesses. Ms. Vogel, I am going to try to pick up where Chairman Foster just left off on compliance testing. Is it fair to say, based on the response you gave, that the right way to think about this is more to look at the outputs of AI as opposed to opening up the hood and trying to understand each individual node and network? Is that the right way to think about it?

Ms. VOGEL. My view is that it should be a balance. I think that, absolutely, the outputs are indicative. They are helpful to look at now because so much of the AI is already deployed, and so we are not at the design stages for so much of the AI in common use, and for that understanding, what the outputs are is important and helpful. I think there are elements of what is under the hood that would be helpful and important to understand, particularly when you are talking about AI used in a pivotal sensitive function. So, I don't think it is one or the other. I do think it is a balance, but no matter what I think, the outputs are very important to be testing and watching.

Mr. GONZALEZ OF OHIO. No, I appreciate that. I think, as I mentioned in my opening statement, it is encouraging that we are seeing some AI algorithms produce better results, significantly better results in some instances, from a bias standpoint. And obviously, the hope is to understand what it is that they are doing right and doing more of that or making that more transparent, and then helping foster a more collaborative innovation environment.

Ms. King, I want to shift to you, and I want to ask about transparency in AI algorithms. Also, as I mentioned in my opening statement, the use of algorithms in social media has had a detrimental effect on young users, which, as a parent, I find extremely problematic. Do you think that more can be done to ensure parents have additional transparency about their child's data being collected by these apps or their own, and how can we strike the right balance and the right line between encouraging innovation, managing problematic algorithms, and providing data sovereignty to users?

Ms. KING. Thank you, sir. As a parent as well, that is the one thing that terrifies me, is my children getting access to these capabilities. And unfortunately, I wish there was one significant answer that could fix it, but it is going to be a constant ever-moving group of things that we have to do. And explainability is significant in that problem because, as Miriam just said, we have to understand what the outputs are, but we have to understand enough about how we are getting there to be able to make informed decisions about whether there is too much data that is being collected or whether there isn't, and there are many ways to do this. One of the most popular is this local and interpretable model agnostic explanation. This was created by the University of Washington to try to see what happens inside, so model agnostic. It should be across models. That is one of many ways to do that.

Another piece to this is, as you just mentioned, that AI can be positive and there are some impressive advances happening right now in synthetic data that can both hopefully correct for some of those historical data biases, but also give just a better picture of the people who are going to be impacted by the system being created. Now, of course, you have to understand what that synthetic data looks like, so you probably should have a wide group of interdisciplinary experts assessing that to make sure you are not missing something. But I think it is a combination of constantly reviewing the outcomes, constantly trying to take at least a sample of explainability across some of the most important, as Europeans are suggesting, high-risk models, and then also assessing kind of what are the new technical capabilities we are developing now that can help address this problem.

Mr. GONZALEZ OF OHIO. Thank you. Shifting to Mr. Cooper for a second with my final minute, I want to ask you about BSA's AI risk management framework included in your testimony. One aspect that seems to be of importance is that a one-size-fits-all framework will not work for small companies and startups. I completely agree. Could you elaborate on why flexibility in any framework is important for fostering innovation?

Mr. COOPER. Sure. Thank you very much. I think it is important to have flexibility in a variety of ways of achieving a desired outcome for a number of reasons, including that not all systems are going to be used for the same purposes. The algorithm and the data that is used to determine what shows our kids watch or what videos our kids watch online is one form of algorithm and one use of AI. But there is also database management, and customer relations management tools, and farmers who use AI in order to improve crop yield, and one set of regulations across-the-board isn't going to be able to be flexible enough to address the range of different use cases for AI.

Mr. GONZALEZ OF OHIO. Yes, thank you. And I also think it is almost always the case that the higher the regulatory burden, the more you entrench incumbents, and the less innovation you have at the startup level as the regulatory burden is just too high to even contemplate a startup. So with that, I thank the witnesses, and I yield back.

Chairman FOSTER. Thank you. The Chair now recognizes Ms. Pressley of Massachusetts for 5 minutes.

Ms. PRESSLEY. Thank you, Chairman Foster, for convening this important hearing, and to our witnesses for joining us here today. Certainly, systemic racial discrimination is widespread in the financial services industry. The damage of redlining, banking deserts, and employment discrimination has never been fully redressed or repaired in America, and all of the data supports those facts. Today, mortgage lenders deny Black applicants at a rate 80 percent higher than White applicants, and payday lenders continue to target low-income people of color, charging 500-percent interest even in the midst of a pandemic. Many believe artificial intelligence presents an opportunity to make the allocation of credit and risk fairer and more inclusive. However, AI technology and machine learning can easily go in the other direction, exacerbating existing bias, and reinforcing bias credit allocation, and making discrimination in lending even harder to prove.

Cases of racial bias in AI are well-documented and have impacted everything from mortgage loans and tenant screening to student loans. The deciding factor between whether the technology has a positive or damaging impact could be its developers.

Ms. Broussard, who is writing the algorithms that are being used to make important financial decisions, like creditworthiness? Do the teams writing these algorithms generally reflect the diversity of people in America?

Ms. BROUSSARD. Generally, these teams do not represent the diversity of people in America. Silicon Valley and its developers tend to be very pale, male, and Yale. Compared to overall private industry, the EEOC found that the high-tech sector employed a larger share of Whites, Asian Americans, and men, and a smaller share of African Americans, Hispanics, and women.

Ms. PRESSLEY. Thank you. In fact, in February 2020, the Financial Services Committee released a report on the diversity of America's largest banks, which found that banks were largely undiversified at all levels and departments. Those data points you offered there support that.

Ms. Broussard, one more question, will this lack of diversity affect AI used by financial institutions? What is the impact?

Ms. BROUSSARD. Absolutely, yes, there is an impact. The problem is that people tend to embed their unconscious biases in the technology that they create. When we have a small and homogeneous group of people creating AI, that AI then gets the collective blind spots of the community of people who are creating the algorithms. So, the more diversity you have in the room when you are creating algorithms, the better the algorithm is going to be for the wide variety of people who live in America.

Ms. PRESSLEY. Thank you, Ms. Broussard. And just to further unpack the impact of that on people's lives, there are many different facets that AI companies developing these technologies really need to consider, from, as we are speaking to here, who is developing the algorithms to the AI's impact on job loss. A recent report from the World Economic Forum predicted that by 2025, the next wave of automation amplified by the pandemic will disrupt 85 million jobs globally.

Ms. Vogel, what role should independent auditors play in helping to assess the human cost and the ethical implications of AI technologies so that both developers and the public can fully understand the ethical impacts these technologies have for actual consumers?

Ms. VOGEL. Thank you for that question. It is a really important point. We do have this growing body of experts—in fact, we have one on this very panel—who do this important work of checking in, of taking the temperature and understanding where these gaps are. I think it is really important that we build our reliance and our infrastructure to support more algorithmic auditing because these are the people who will tell us if the AI doing what we expect it to. Are we discriminating? Are we creating opportunity? For whom will this fail, and how do we create more opportunity through our algorithms?

Ms. PRESSLEY. Thank you. I agree. Frequent and independent audits are critical. AI-supported recommendations in the financial services industry directly impact people's lives and economic opportunities, and yet the algorithms used are trained on data that is rife with imbalance and discrimination. So as we do the work deliberately to enact long-overdue economic justice, we can't allow the AI industry to create new problems and to compound these already persistent and deeply-embedded inequities. Thank you, and I yield back.

Chairman FOSTER. Thank you.

The Chair now recognizes Mr. Loudermilk from Georgia for 5 minutes.

Mr. LOUDERMILK. Thank you, Mr. Chairman. One important thing to keep in mind as we discuss AI is the types of bias we need to eliminate and the types of bias that we actually want to keep. Sometimes, the main purpose of an algorithm is to be biased. For example, in loan underwriting, algorithms are generally used to distinguish between who can pay back a loan and who is not able to pay back a loan. With that in mind, we must work toward eliminating the types of bias that have no place in our financial system, such as the bias based on race or gender or any other factor like that.

One important way of doing that is when an algorithm is being built, there should be a thorough record-keeping of everything that is added to the algorithm. That way, if bias is suspected, companies and regulators can see everything that went into the algorithm and see where the bias may be coming from. I think this would help make it where algorithms are not a black box and where the outputs cannot be explained, but you would have a record where you could see where the problems may be.

Mr. Cooper, your organization's framework for AI best practices recommends maintaining records of the data that is used to train AI models. I agree with that. Expanding on that, do you believe that maintaining thorough records of all of the inputs used to build an algorithm can be useful for identifying the source of any unwanted bias?

Mr. COOPER. Thank you very much, Congressman. Yes, I think it goes even beyond what the data is. I think that there is a whole set of considerations that companies need to go through to figure out whether there is a high risk that the AI system, as it is intended to be deployed or as it is being deployed, may have consequential impacts on people. And the decision-making about what those risks are and what the right mitigation practices are, how the data was tested, what historical biases may or may not be present in them, keeping a record of that as part of a risk management framework, can be both useful in order to make sure that companies are not putting systems out into the world or using systems that are going to lead to discriminatory results. But it could also be useful, as you say, after the fact, if there is a problem, to go back and audit and find out why it happened and make sure it doesn't happen again.

Chairman FOSTER. Representative Loudermilk, I believe you are muted.

Mr. LOUDERMILK. I don't know why it is cutting off like that. Can you hear me now?

Chairman FOSTER. Yes, we can.

Mr. LOUDERMILK. Okay.

Chairman FOSTER. And feel free to exceed your time by 40 seconds.

Mr. LOUDERMILK. Thank you. Mr. Yong, in your testimony, you discuss the importance of accountability and transparency in AI. Can maintaining records when algorithms are being built help achieve those goals?

Mr. YONG. Yes. Accountability is very important, especially when it comes to AI, and without record-keeping, there is no transparency. In our testimony, we mentioned that transparency is a prerequisite to enabling financial institutions to meet the other general AI governance principles. And if the AI model is not transparent, then it is very difficult to assess whether it is reliable, whether it is sound, and whether there is bias involved. So definitely, record-keeping is a prerequisite to meeting this accountability and general principles.

Mr. LOUDERMILK. Thank you for that. Ms. King, you have written that policymakers must govern AI in a way that is flexible enough to adapt when technology inevitably changes, wand we know that it continually changes in today's environment. I agree with that, and I believe that is needed to have an environment that fosters innovation. With that in mind, how can policymakers ensure that AI governance remains flexible, but robust, at the same time?

Ms. KING. Thank you, sir, for that question, and I think it is all about having a set of goals that are measurable and achievable. One of them is, how can you explain these systems, and, again, the complexity here is really because these systems cross so many sectors. Yours obviously is financial services, so you have some very specific use cases to identify, which is helpful, but you need to be able to explain those specific use cases. You need to ask a lot of questions, and those questions will change, too, but the big ones are why was it developed. What are the [inaudible]? How does it possibly fail because it is the unexpected failure that is really a lot of the problem here.

And then again, how can we correct those errors and report them? So if you can kind of have those four ways of addressing this challenge and work with companies, and you work with both governments who are buying this and the companies who are producing this to have sort of the four methods of regularly checking that you are getting, you are producing what you want, you are getting what you want out of it, and that it is not discriminatory, then I think that is a flexible way to move forward.

And I think the sandbox concept that Ms. Broussard has suggested is also very helpful, because while records are great and it is easy for us to say, let's keep records here, if you have ever taken a look at the code behind some of these systems and how often it changes as you shift weights, it gets pretty complicated pretty quickly. So, the more you can have these kinds of places where companies and organizations can feel safe testing is going to be critical going forward well.

Mr. LOUDERMILK. Thank you. And, Mr. Chairman, I yield back the balance of my time.

Chairman FOSTER. Thank you,.

The Chair now recognizes Representative Casten for 5 minutes.

Mr. CASTEN. Thank you so much. I think Mr. Loudermilk really hit it on the head with the transparency question, and I want to follow on that, but I want to specifically get to the auditability issue, and I think you alluded to this, Ms. King. It is one thing to be able to see the code. It is something else completely to be able to understand the code. And I say this as someone who, before I came to Congress, ran a utility. And my biggest risk was predicting revenue—did it vary with the weather, did it vary with economic conditions—and I built a genetic algorithm that figured all that stuff out. I have no idea how it worked, but it was amazingly effective and it made our investors much happier because we could predict our revenue.

That is trivial. It is not at all implausible for me to imagine that we get to a world where an investment fund has figured out from looking at global data that there is about to be a massive human rights abuse committed, and it is shorting the affected businesses and properties, right? That would be deeply unethical, and if we understood it, it would be a problem, but it is totally possible that we could never actually understand that and saying that is what it is doing.

So my question is, and I think all of you can answer this, but I am going to start with Ms. King, just because I see you nodding so vociferously, what is the best regulatory practice for ensuring that these algorithms remain auditable, and ensuring that they apply to everyone in the system, because presumably, as soon as some subset of people agree to have auditable algorithms, people who violate that might have an investing edge, whether that is a bad actor in our country or a foreign actor who wishes us harm. What is that standard, both domestically and internationally? How would you recommend we think about that?

Ms. KING. Thank you, sir. I will take a quick stab at it, and I am going to use a hypothetical because I think it is always helpful to have it. Yours is very complicated, and I am not going to try and explain your very impressive example. At the Wilson Center, in one of our trainings, we use a supply chain prediction model. Will it or won't it arrive? Will the USPS deliver a package on time? And you have a series of variables. You have product wait. You have the month that it was ordered. You have things that you probably, as a consumer, wouldn't think matter, but about 10 different variables. And as you play with the model and you change the variables, you change the weights—how much weight do we give to a particular variable or not—you understand more why the prediction you get comes out, and then you can kind of take that and you can go back through the system and check it.

I would say you need a couple of standards. One is not going to work, unfortunately, but a couple of standards that have that sort of ability to use a couple of methods, probably a model agnostic method, if it is possible, to go back and just understand, at least at a strategic level. You may not be able, as you know very well, to go and explain the whole thing, but a confidence level and then explainability that you could achieve. So, you are looking for some sort of confidence trust level and some sort of agnostic model verification, and you are also looking to make sure as you are going through that process, that if you are going to have regulators as part of this conversation, you have a number of regulators across sectors. As your example points out, you can't just have financial services regulators. You are going to have to have others from other parts of the government at the table because of these unexpected outcomes.

Mr. CASTEN. If I could, though, that approach you described works where there is a finite number of known inputs. If you are using sort of a neural network model that, for all practical purposes, has an infinite number of inputs, I don't know how you audit that at some level of complexity. To follow on from that, and I am sure this varies market to market, is there any good analysis? Is there a percentage of algorithmic trading or algorithmic investing, wherever it sits, that we really don't want to have more than X percent because now the algorithms are responding to algorithms? Is there a robust mathematical way to think about that? And maybe it is different for housing credit decisions than it is for equities investments or something else. But is there some robust way to think about that so that we don't sort of unwittingly introduce too much volatility into the system? And if any of the other witnesses want to chime in on this, you would be welcomed for your thoughts as well.

Ms. VOGEL. I can speak to auditing algorithms. What we want to do is, we don't want to think about auditing all algorithms to the same standard. We want to think about auditing algorithms in context because the context matters a lot. So, we do need to keep track of inputs. We do need explainability. We do need to enforce real-world laws inside algorithms. We do you need to be aware of bias in, bias out. And so to your point about thresholds, the acceptable thresholds would be determined based on the context.

Mr. CASTEN. I see I am out of time, but I would welcome further thoughts offline from any of the witnesses, and I yield back.

Chairman FOSTER. Thank you.

The Chair now recognizes Ms. Adams from North Carolina for 5 minutes.

Ms. ADAMS. Thank you, Chairman Foster. And thank you, Ranking Member Gonzalez and Chairwoman Waters, for this hearing today. And to our witnesses, thank you for your testimony.

Professor Broussard, in your testimony, you noted that all historical data sets have bias, and that AI needs to be regulated as soon as possible because it has all of the flaws of any human process plus more. You also cite in your testimony the potential impact of bias in AI to students and consumers of lower socioeconomic status, such as when the International Baccalaureate used AI to assign grades to students, to their detriment. So, building off of what my colleague, Ms. Pressley, discussed, would you tell us more about what happened in these scenarios?

Ms. BROUSSARD. Sure. Thank you for that question. The International Baccalaureate (IB) example is a situation where, because of the pandemic, the International Baccalaureate exams were canceled, and the IB decided to use an algorithm to assign imaginary grades to real students, which had disastrous consequences, because the inputs to the algorithm were things like a school's performance in the past. We know that the economic divide is particularly profound when it comes to America's schools, and so the students at the poor schools were predicted to do poorly, and the students at the rich schools were predicted to do well. We have a racial divide there. Who are the students at poor schools? They are mostly Black and Brown students. Who are the students at rich schools? Well, they are mostly White students. So, the algorithm made very predictable decisions that disadvantaged Black and Brown and poor students. This is what happens most of the time with algorithmic decisions.

Ms. ADAMS. Would you explain what algorithmic auditing is, and how we can encourage public and private entities to adapt it as a best practice?

Ms. BROUSSARD. Thank you. Yes. Algorithmic auditing, as I mentioned in my testimony, is something that I do with a company called O'Neil Risk Consulting and Algorithmic Auditing, Inc. (ORCAA). What we do is we look at an algorithm and we ask, who could this algorithm negatively affect, and we look at the inputs to the algorithm. We do look at the code. We act as an information fiduciary, so we keep everything extremely private. We look at the outputs and we do mathematical and statistical analysis as necessary in order to figure out what is going on in the algorithm. Once you actually figure out where the algorithm is going wrong, you can fix it, but in a lot of industries now, people are pretending that there is nothing wrong. For example, Ms. Vogel mentioned before the Optum case. There is also the case of the Apple card, where a man was offered a credit limit that was about 10 times higher than his wife, even though they shared all of their finances.

Companies are pretending that they don't collect information like race in order to make decisions. But, on the other hand, if you are using a factor, like a ZIP Code, that is an input to your algorithm, then, actually we have enough residential segregation in the United States that if you are using a ZIP Code, you are actually using race as a proxy. So, there are—

Ms. ADAMS. Thank you.

Ms. BROUSSARD. Thank you.

Ms. ADAMS. I want to move on, if I can, quickly.

Ms. BROUSSARD. Sure.

Ms. ADAMS. Ms. Vogel, I was happy to see that part of your recommendations related to diversifying the AI field, including supporting Historically Black Colleges and Universities (HBCUs). Specifically, what shoud Congress be doing to ensure that HBCU and Minority Serving Institution (MSI) students are able to participate in the AI revolution that is currently underway?

Ms. VOGEL. Thank you for that question. We strongly believe that we need AI to be built by and for a broader cross-section of the population, both so that more can benefit from the AI, so that more can benefit from the economic support that comes from it, but also so that our AI is better. So, we need to make sure that we support HBCUs and MSIs to ensure that their students are part of this current AI revolution that is undeway. We know that HBCUs produce nearly 20 percent of all Black graduates, 25 percent of Black graduates who earned degrees in the disciplines of STEM technology, science, engineering, and math, and we need to make sure that we have all hands on deck. We can't afford to not bring all of these students into the AI revolution. Industry is depending on their participation.

Ms. ADAMS. Thank you, ma'am. I think I am out of time. Mr. Chairman, I am going to yield back, but thank you very much for your response.

Chairman FOSTER. Thank you.

The Chair now recognizes Mr. Auchincloss of Massachusetts for 5 minutes.

Mr. AUCHINCLOSS. Thank you, Mr. Chairman. I would like to talk about two specific applications of algorithms that have been and are front and center these days, and really invite the panel to weigh in on one or both of them. The first is the use of algorithms in hiring. A number of organizations, some from the center-left, some from the left, and some from the center-right, have all converged that there are somewhere between 25 to 30 million "hidden workers" in the United States, people who could be employed, who, under the right conditions, want to be employed. And yet, we are not tapping into their productivity, and they are not getting to realize their fullest aspirations.

That is obviously a multifaceted problem, but one element of it is algorithms that some of the biggest companies are using, something like 75 to 80 percent of Fortune 100s, for example, in how they sort resumes that get put forward. They are screening out resumes that have discontinuity in employment. They are screening out resumes of formerly-incarcerated individuals. They are screening out resumes that don't have a college degree, even for jobs that don't require a college degree. I welcome input from the panel on this first application, kind of the state of play right now in these resume-screening algorithms, and what can be done to improve them, and whether they have any role at all going forward?

Ms. BROUSSARD. I can offer that my colleague, Hilke Schellmann, has been writing about these topics, and has done some really excellent work in the MIT Technology Review, that is an in-depth review on what is going on with hiring algorithms.

Mr. COOPER. Yes. This is one of the reasons why we need a risk management framework for when there is going to be an AI system that has a highly-consequential impact on somebody's lifec, so making sure that there is a thought process that is auditable about what the factors are that are considered in determining what resume is going to go where is a good example of something where we need to make sure that there is a thought-through and documented impact assessment, and then steps taken to mitigate risk.

Mr. AUCHINCLOSS. And I would just add also that this should be a triple-line win for everybody. Companies don't want to be screening out high-quality workers for esoteric reasons. They are struggling for employees, as we speak. And as a society, we want people to be working and contributing, and people themselves want meaningful work. So, I would hope that this can be an area of actual bipartisan work going forward on how we encourage the private sector to be more thoughtful about their use of these hiring algorithms and the other elements of this challenge of hidden workers.

The second area that I want to dig into and invite the panel to speak on is about what many of us have been reading about these last 2 weeks, which is Facebook's algorithm. The whistleblower addressing the Senate exposited that while she did not think Section 230 should be revised for user-generated content, she did think that Facebook's algorithm itself should be subject to liability laws. And I would welcome input from any of the panelists here about how that might be applicable in terms of Facebook, in particular, but really any social media's algorithm, whether that should be subject to regulation itself?

Mr. COOPER. I am happy to jump in again. We don't represent Facebook, but I would say that I think it is important to make sure that where you have particular high risk in the way an algorithm is working—in this case, feeding certain videos or certain social media feeds to certain people, particularly where it has to do with children—is a high risk, and we need to make sure that the decision-making process is appropriate. And there is a combination of a regulatory aspect of that and also just good practices internally to make sure that there is organizational accountability so that when decisions are made, that there is somebody at a senior level who signs off on those decisions, and that there is documentation of why certain choices were made.

Mr. AUCHINCLOSS. Yes. I can see why it would be challenging to try to unpick liability for an algorithm that was put into place, how you can draw causality directly, and yet part of me thinks that we have to answer that question. We have to wrestle with that problem because, otherwise, we are going to be in a place where I think organizations will be distancing themselves from accountability instead of embracing it by being able to point towards these black box algorithms and say that they are just part of part of their toolkit, and you can never pay cause and effect. I think we need to reject that explanation and hold companies liable for the algorithms that they choose to use.

I yield back my time, Mr. Chairman.

Chairman FOSTER. Thank you, and I would like to thank our witnesses for their testimony today.

The Chair notes that some Members may have additional questions for these witnesses, 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 now adjourned.

[Whereupon, at 1:16 p.m., the hearing was adjourned.]

APPENDIX

October 13, 2021

Statement by
Meredith Broussard
Associate Professor, New York University
Research Director, NYU Alliance for Public Interest Technology
before the
Task Force on Artificial Intelligence
of the
Committee on Financial Services
U.S. House of Representatives

Representative Foster, members of the Task Force, thank you for hosting this important hearing on ethics in artificial intelligence, and for giving me the opportunity to submit this testimony. My name is Meredith Broussard and I am an associate professor at the Arthur L. Carter Journalism Institute of New York University, the research director at the NYU Alliance for Public Interest Technology, and an affiliate of the NYU Center for Data Science. I'm also the author of the book, *Artificial Unintelligence: How Computers Misunderstand the World,* which has been widely adopted as a text in AI ethics courses. I began my career as a computer scientist at AT&T Bell Labs and the MIT Media Lab before turning to journalism, where I now teach investigative journalism using data and code. As part of my research, I create artificial intelligence for investigative reporting, and I do a lot of science communication work around computational literacy in order to empower people to understand the algorithms that are increasingly used to make decisions on our behalf. I also consult on algorithmic audits of commercial systems; I am working on developing a regulatory sandbox in order to audit AI systems for legal compliance; and I founded a summer program for early and mid-career scholars called the NYU Institute for Public Interest Technology.

In this testimony, I'm going to explore a practical vision for regulating artificial intelligence that builds on the wide-ranging testimony that has already been presented before this Committee. I'll do a few things:

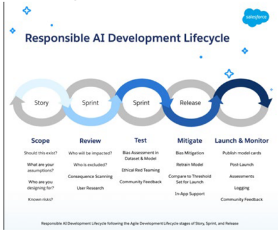
- Explain what AI is and isn't
- Talk about discrimination by default
- Talk about algorithmic auditing
- Explain that some frameworks for AI ethics exist, and note how they can be integrated into business processes
- Talk about regulatory sandboxes, which are a promising development for auditing and compliance

The first thing I want to say is that AI is not what we see in Hollywood depictions. There is no robot apocalypse coming, there is no Singularity, we do not need to prepare for artificial general intelligence (AGI) because these things are imaginary. What is real is that AI is math. General AI is the Hollywood sci-fi version, and it is entirely imaginary. Narrow AI is what we have, and it is very complicated and beautiful math. It is math that is computed on machines. I say this in order to underscore the fact that computing is a terrestrial process; it does not take place in a literal cloud. The cloud is someone else's computer. The process of "doing AI" is something that humans do with machines. As a sub-field of computer science, AI itself has many sub-fields. Machine learning is currently the most popular. Machine learning is a sub-field of AI, the same way that algebra is a sub-field of mathematics. However, the terms "AI" and "machine learning" tend to be used interchangeably today. Both are poorly-chosen terms, because they suggest there is a brain, or sentience, inside the computer. There is not. When we do machine learning, we take a large set of historical data, and instruct the computer to create a model based on patterns and values in that dataset. The model can then be used to predict or make decisions based on past data. The more data you put in, the more precise your predictions will become. Computer scientists refer to this as the “unreasonable effectiveness of data.” However, all historical datasets have bias. For example: if you feed in data on who has gotten a mortgage in the past in the United States, and ask the computer to make similar decisions in the future, you will get an AI that offers mortgages to more white people than BIPOC people.

AI is a terrestrial process, and it needs to be regulated ASAP because it has all of the flaws of any human process, plus some. The previous recommendations offered before this committee have offered detail on additional factors such as identity verification and cybersecurity, which are of course important parts of the landscape in regulating AI.

My current regulatory vision begins with frameworks, high-level governance models that guide a company’s use of AI and data. A company can make sure its frameworks are implemented by performing regular algorithmic audits, ideally using a regulatory sandbox. The process could be monitored by regulators using tools we already have, namely compliance processes inside existing regulatory agencies.

Many frameworks for AI ethics and bias detection exist. Salesforce’s AI ethics lead, Kathy Baxter, has helpfully gathered many of them online. I like the ethical AI checklist developed by Equal AI, an organization that is led by Miriam Vogel, who is also testifying today. NIST is developing an Artificial Intelligence Risk Management Framework intended for voluntary use and to improve the ability to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems. Framework concepts can be integrated into normal business processes. The Salesforce site that I mentioned also includes a diagram showing how bias detection and auditing can be integrated into agile software development methods, as follows:



Responsible AI Development Lifecycle following the Agile Development Lifecycle stages of Story, Sprint, and Release

Every company needs a framework for AI governance, and needs to implement the concepts. The next question becomes: inside a company, which AI needs to be regulated and monitored? This depends on the user and the context. Automated license plate readers used at toll booths by the local department of transportation, with data stored for only a short time, is a reasonable use of AI. Automated license plate readers used by police as dragnet surveillance, with the data stored indefinitely, is an unreasonable use of AI.

The EU’s proposed AI regulation calls for categorizing AI into high and low risk. A low-risk use of facial recognition might be using facial recognition to unlock your phone. This is fairly inoffensive, and there is a fallback (a PIN code) for when the facial recognition technology fails. A high-risk use of facial recognition might be the police using facial recognition on real-time surveillance video feeds. Facial recognition technology has been shown to consistently mis-identify people with darker skin; people of color are at a high risk of being harmed by facial recognition when it is used in policing. High-risk AI would need to be registered and regularly audited to ensure it is not harming citizens. This EU regulation is a good start, and I would recommend the US adopts a similar strategy of characterizing AI as the high-risk and low-risk, and regulating the high-risk uses in each industry.

After deciding which AI gets regulated, it is necessary to look for specific kinds of bias. The process for uncovering algorithmic bias is called algorithmic auditing. O'Neil Risk Consulting and Algorithmic Auditing (ORCAA), a company I consult with, performs bespoke algorithmic audits that are tailored precisely to a company's needs. ORCAA audits algorithms in context, asking how an algorithm might fail and for whom. This is a way of identifying how an algorithm might be racist or sexist or ableist or might discriminate illegally—and once we identify the problem, it can be addressed, or the algorithm can be discarded. Software like Parity or Aequitas or Fairness 360 can evaluate algorithms for one of 21 known kinds of mathematical fairness.

The proposed EU legislation calls for the use of a regulatory sandbox, which I am particularly enthusiastic about. A regulatory sandbox is a protected environment where companies can test their algorithms for bias. If and when the bias is discovered, they can then address the issue in their code and re-run the test until they are in compliance with acceptable thresholds. Currently, heavily regulated industries like insurance make the claim that they are not collecting race data, and thus their AI can't be biased. Other factors like zip codes operate as proxies for race, however. If an AI uses zip code in order to determine the price of an insurance policy, it is using race as a factor, which is a problem. Using a regulatory sandbox would allow an insurance company to see if they are inadvertently using a protected characteristic to make a coverage decision, and would allow them time to address the issue instead of pretending it does not exist. I'm currently working with ORCAA to develop a regulatory sandbox prototype. In our version, regulators would also have a limited view inside the sandbox, to see that companies are auditing their algorithms for bias and fixing the problems that they find. Our concept also allows regulators to see reports showing algorithmic audit results without the companies revealing any trade secrets.

An open secret in the AI world is, everyone knows these systems discriminate. Any conversation about robot apocalypse is a deliberate distraction from the harms that AI systems are causing today, right now. AI is preventing people from getting mortgages: a recent investigation by The Markup found that nationally, loan applicants of color were 40%–80% more likely to be turned down by mortgage-approval algorithms, as compared to their White counterparts. In certain metro areas, the disparity was greater than 250%. When the International Baccalaureate used AI to assign student grades during the pandemic, high achieving low-income students received terrible grades, which prevented them from getting college credits that would allow them to graduate early and incur less student loan debt. AI is used to generate secret predictive consumer scores, like health risk scores, consumer prominence scores, identity and fraud scores, or summarized credit statistics. It is likely that BIPOC people are systematically disadvantaged by most of these scoring systems. AI is used in so-called predictive policing. One particularly egregious example is found in Pasco County, Florida, where the Sheriff's office used AI to generate a list of people who were predicted to be at risk of becoming criminals in the future, though they had done nothing wrong. The police then pre-emptively harassed the people on this list, which included students who were identified based on their educational records. AI is not a magic bullet; it may seem to solve certain business problems, but it inevitably causes new problems and has unintended consequences.

A useful frame is found in Ruha Benjamin's book *Race After Technology*, in which she argues that automated systems discriminate by default. If we adopt this vision, it becomes dramatically easier to spot discrimination and bias inside AI systems. It is not a question of whether, but a matter of looking for the obvious.

Companies should be evaluating the potential benefits, harms, and greater implications of their AI technology, rather than enthusiastically adopting every new technology in a mad scramble to emulate Big Tech firms. I'd like to encourage a space for technology refusal, normalizing and rewarding the firms who refuse to use AI systems that are biased or discriminatory.

I've laid out a vision here. In my vision, companies adopt meaningful AI ethics frameworks; algorithmic audits are seamlessly integrated into business processes; we have a comprehensive regulatory policy that mandates algorithmic auditing for AI; we have regulators who are trained to spot bias in AI systems; and a new kind of technology will have been developed to facilitate and monitor the process. The final piece is education. Companies need to educate their workers in AI. This might mean calling on groups like the NYU Alliance for Public Interest Technology for professional development. It might mean executives reading books like *Artificial Unintelligence*, or *Algorithms of Oppression* by Safiya Noble, to get better informed about the limits of AI and how bias operates inside sociotechnical systems. Education will also be needed around new AI compliance measures, so that people understand better how bias manifests and how to detect it and address it inside AI systems.

The technology to achieve dramatically better algorithmic insight already exists. Various mathematical definitions of fairness exist. Parity and Aequitas and Fairness 360 are all platforms for algorithmic auditing and bias detection. ORCAA's regulatory sandbox builds on their excellent work. All of the necessary pieces have been made incarnate through the hard work and creativity of scholars, activists, and concerned parties. The missing link is the policy mandate. I would welcome the opportunity to talk more about AI regulatory policy. Thank you for the opportunity to testify today on this important topic.

Truth in Testimony Disclosure Form

In accordance with Rule XI, clause 2(g)(5)* of the *Rules of the House of Representatives*, witnesses are asked to disclose the following information. Please complete this form electronically by filling in the provided blanks.

Committee: US House Committee on Financial Services

Subcommittee: Task Force on Artificial Intelligence

Hearing Date: 10/13/2021

Hearing Title :

Beyond I, Robot: Ethics, Artificial Intelligence, and the Digital Age

Witness Name: Aaron Cooper

Position/Title: Vice President for Global Policy

Witness Type: ◎ Governmental ○ Non-governmental

Are you representing yourself or an organization? ◎ Self ○ Organization

If you are representing an organization, please list what entity or entities you are representing:

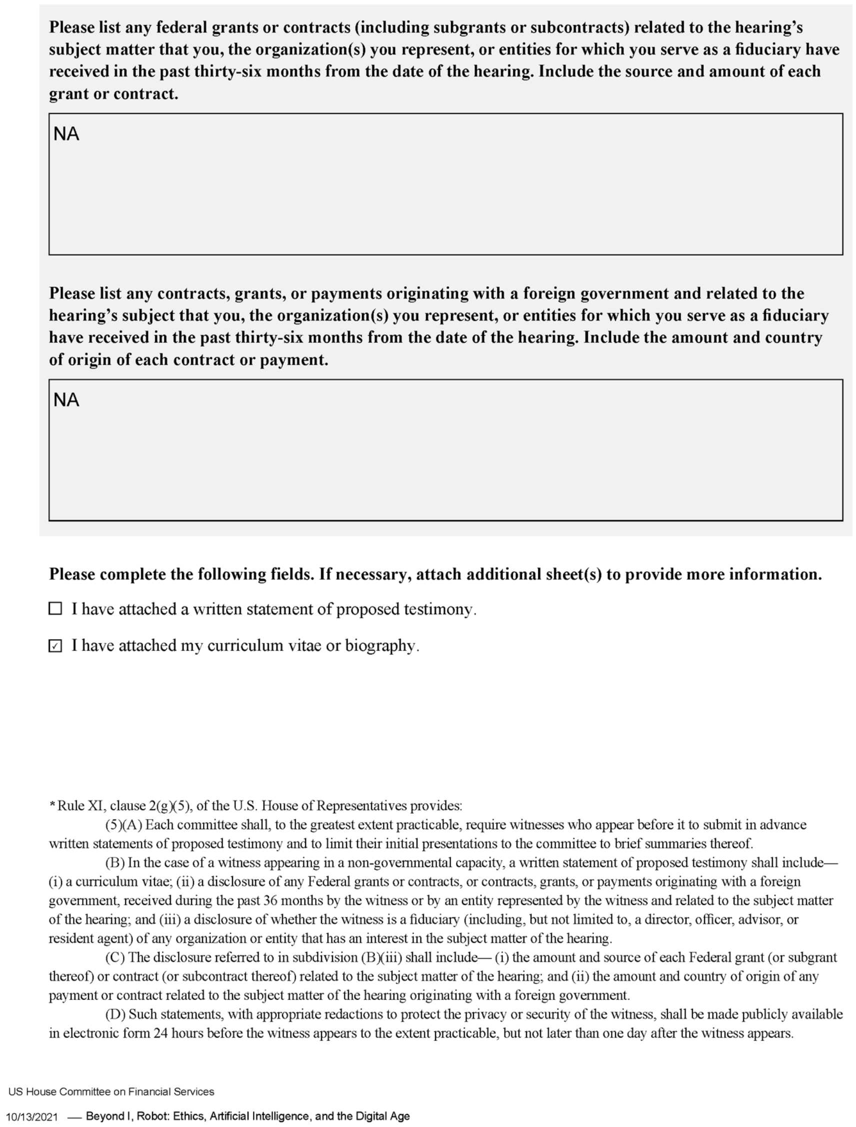
BSA | The Software Alliance, Inc.

FOR WITNESSES APPEARING IN A NON-GOVERNMENTAL CAPACITY

Please complete the following fields. If necessary, attach additional sheet(s) to provide more information.

Are you a fiduciary—including, but not limited to, a director, officer, advisor, or resident agent—of any organization or entity that has an interest in the subject matter of the hearing? If so, please list the name of the organization(s) or entities.

No



Please list any federal grants or contracts (including subgrants or subcontracts) related to the hearing's subject matter that you, the organization(s) you represent, or entities for which you serve as a fiduciary have received in the past thirty-six months from the date of the hearing. Include the source and amount of each grant or contract.

NA

Please list any contracts, grants, or payments originating with a foreign government and related to the hearing's subject that you, the organization(s) you represent, or entities for which you serve as a fiduciary have received in the past thirty-six months from the date of the hearing. Include the amount and country of origin of each contract or payment.

NA

Please complete the following fields. If necessary, attach additional sheet(s) to provide more information.

□ I have attached a written statement of proposed testimony.

☑ I have attached my curriculum vitae or biography.

*Rule XI, clause 2(g)(5), of the U.S. House of Representatives provides:

(5)(A) Each committee shall, to the greatest extent practicable, require witnesses who appear before it to submit in advance written statements of proposed testimony and to limit their initial presentations to the committee to brief summaries thereof.

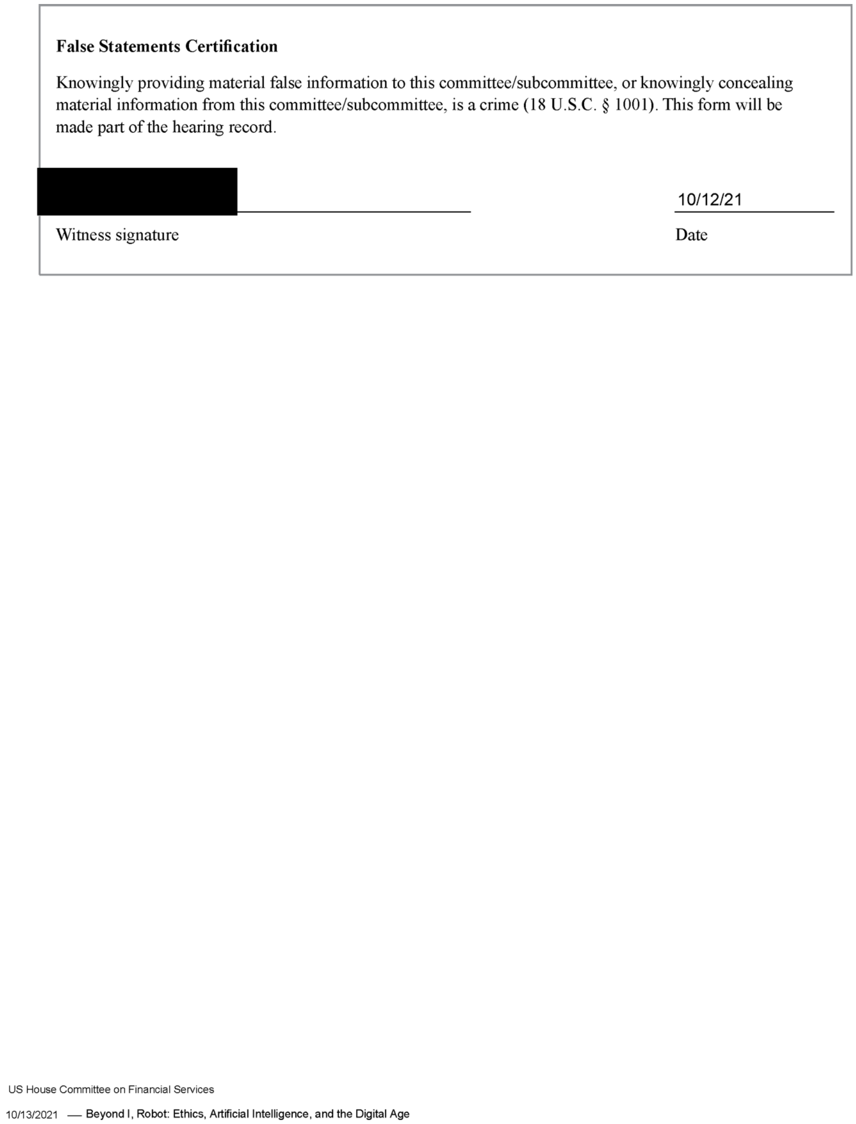
(B) In the case of a witness appearing in a non-governmental capacity, a written statement of proposed testimony shall include— (i) a curriculum vitae; (ii) a disclosure of any Federal grants or contracts, or contracts, grants, or payments originating with a foreign government, received during the past 36 months by the witness or by an entity represented by the witness and related to the subject matter of the hearing; and (iii) a disclosure of whether the witness is a fiduciary (including, but not limited to, a director, officer, advisor, or resident agent) of any organization or entity that has an interest in the subject matter of the hearing.

(C) The disclosure referred to in subdivision (B)(iii) shall include— (i) the amount and source of each Federal grant (or subgrant thereof) or contract (or subcontract thereof) related to the subject matter of the hearing; and (ii) the amount and country of origin of any payment or contract related to the subject matter of the hearing originating with a foreign government.

(D) Such statements, with appropriate redactions to protect the privacy or security of the witness, shall be made publicly available in electronic form 24 hours before the witness appears to the extent practicable, but not later than one day after the witness appears.

US House Committee on Financial Services

10/13/2021 — Beyond I, Robot: Ethics, Artificial Intelligence, and the Digital Age



False Statements Certification

Knowingly providing material false information to this committee/subcommittee, or knowingly concealing material information from this committee/subcommittee, is a crime (18 U.S.C. § 1001). This form will be made part of the hearing record.

[redacted] ______________________ 10/12/21

Witness signature Date

Aaron Cooper
Vice President, Global Policy
BSA | the Software Alliance, Inc.

Aaron Cooper serves as Vice President, Global Policy, of BSA | The Software Alliance. In this role, Cooper leads BSA's global policy team and contributes to the advancement of BSA members' policy priorities around the world that affect the development of emerging technologies, including data privacy, artificial intelligence, cybersecurity, intellectual property, and trade. Cooper joined BSA in February 2016 as Vice President, Strategic Policy Initiatives.

Cooper previously served as the Chief Counsel for Intellectual Property and Antitrust Law for Chairman Patrick Leahy on the U.S. Senate Judiciary Committee. Most recently, Cooper was of counsel at Covington and Burling, where he provided strategic counseling and policy advice on a broad range of technology issues. Cooper has also served as Legal Counsel to Senator Paul Sarbanes.

Cooper has testified before Congress and is a frequent speaker on data privacy and security, intellectual property, trade, and other issues important to the software industry.

Cooper is a graduate of Princeton University and Vanderbilt Law School. He clerked for Judge Gerald Tjoflat on the U.S. Court of Appeals for the Eleventh Circuit.

Hearing on

“Beyond I, Robot: Ethics, Artificial Intelligence, and the Digital Age”

United States House of Representatives Committee on Financial Services
Task Force on Artificial Intelligence

October 13, 2021, at 12:00 p.m.

Testimony of Aaron Cooper
Vice President, Global Policy
BSA | The Software Alliance

Testimony of Aaron Cooper
Vice President, Global Policy, BSA | The Software Alliance
Hearing on "Beyond I, Robot: Ethics, Artificial Intelligence, and the Digital Age"

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

October 13, 2021

Good afternoon Chairman Foster, Ranking Member Gonzalez, and members of the AI Task Force. My name is Aaron Cooper. I am Vice President of Global Policy for BSA | The Software Alliance (BSA).

BSA is the leading advocate for the global software industry.[1] Our members are at the forefront of developing cutting-edge, data-driven services that have a significant impact on US job creation and growing the global economy. I commend the Task Force for convening today's important hearing, and I thank you for the opportunity to testify.

Enterprise software services, including artificial intelligence (AI) are accelerating digital transformation in every sector of the economy. Artificial intelligence is not just about robots, self-driving vehicles, or social media. It can be used by businesses of all sizes to improve their competitiveness, enhance their value proposition, and increase their capacity to make data-informed decisions.

BSA represents the perspective of the enterprise software companies that help make this possible. Our members create the technology products and services that help other businesses innovate and grow. In that capacity, BSA members are on the leading edge of providing businesses in every sector of the economy with the trusted tools they need to leverage the benefits of AI.

The promise that AI may one day impact every industry is quickly turning into a commercial reality and driving the digital transformation. For instance, Autodesk brings the power of AI to industrial design, helping American manufacturers improve the performance of their products while reducing their costs and environmental impact. In one recent collaboration, Autodesk worked closely with engineers at General Motors to explore how AI-enabled generative design could help the company optimize its manufacturing processes.[2] As an initial proof-of-concept, the two companies set out to improve GM's approach to designing and manufacturing the brackets that secure seatbelts and seats to a car's floor. The partnership yielded immediate benefits, enabling GM to identify a new design that is 40 percent lighter and 20 percent stronger than its previous approach.

[1] BSA's members include: Adobe, Atlassian, Autodesk, Bentley Systems, Box, CNC/Mastercam, DocuSign, IBM, Informatica, MathWorks, Microsoft, Okta, Oracle, PTC, Salesforce, ServiceNow, Siemens Industry Software Inc., Splunk, Trend Micro, Trimble Solutions Corporation, Twilio, Workday, Zendesk, and Zoom Video Communications, Inc.

[2] General Motors | Generative Design in Car Manufacturing | Autodesk

Splunk is helping the financial services sector leverage AI to take a bite out of the more than $40 billion that is lost to fraudulent transactions each year. Splunk's software powers a suite of enterprise fraud management capabilities that allow banks to identify transaction anomalies in real time, reduce the frequency of false positives, and better protect consumers from identity theft.[3]

While the adoption of AI can unquestionably be a force for good, it can also create real risks if not developed and deployed responsibly. We commend the Task Force for convening today's hearing to examine the role that frameworks for ethical AI can play in ensuring the responsible use of this technology. This is an area of particular focus for BSA and our member companies are leaders when it comes to responsible AI practices.[4] We recently produced a detailed framework that sets forth a risk management approach for confronting concerns about bias. As the Task Force explores the use of these tools, we offer our perspective on how they can be used to address the risk of bias, which we hope will also inform the broader conversation at the hearing today.

As this Task Force is aware, the data-driven nature of AI makes it susceptible to unintentional bias. Because AI is trained on data from the past, there is a risk that AI systems may replicate and potentially further entrench historical inequities. As AI is integrated into business processes that can have consequential impacts on people's lives, there is a risk that "biased" systems will perform less accurately or unfairly disadvantage members of historically marginalized communities.

For BSA members, earning trust and confidence in the AI and other software they develop is crucial, so identifying ways to reduce the risk of bias is a priority. BSA therefore set out to develop real, credible, and actionable steps to guard against the potential of AI systems producing unintended disparate impacts. The resulting framework – Confronting Bias: BSA's Framework to Build Trust in AI – was released in June and is built on a vast body of research and informed by the experience of leading AI developers.[5]

The Framework outlines a lifecycle-based approach for performing impact assessments to identify risks of AI bias and corresponding best practices for mitigating those risks. The foundation of the Framework is its detailed methodology for performing impact assessments that help ensure that critical decisions are documented and that an organization's product development team, its compliance personnel, and senior leadership are aligned on the appropriate steps for mitigating risks of bias when they are identified. The Framework is intended to scale with risk and recognizes that inherently low-risk systems—for example, a system used to predict the type of fonts being used on a document—may not require a full impact assessment. But for systems that pose heightened risks, a robust impact assessment is essential to mitigating potential harms.

[3] Detecting Credit Card Fraud Using SMLE | Splunk; Splunk at TransUnion | Splunk

[4] See, e.g., Adobe - Adobe unveils new AI ethics principles as part of commitment to responsible digital citizenship; IBM - 3 lessons from IBM on designing ethical AI technology | World Economic Forum (weforum.org); Microsoft - Our approach to responsible AI at Microsoft; Salesforce - Salesforce Debuts AI Ethics Model: How Ethical Practices Further Responsible Artificial Intelligence - Salesforce News; Workday - Building Trust in AI and ML Through Principles, Practice, and Policy (workday.com).

[5] Confronting Bias: BSA's Framework to Build Trust in AI

The Framework is ultimately a playbook that organizations can use to enhance trust in their AI systems through risk management processes that promote fairness, transparency, and accountability, three of the key principles for responsible and ethical AI. The full Framework, with more than 50 actionable diagnostic statements, is attached to my testimony and can be found at ai.bsa.org. Below, I share a few key insights from the Framework.

Overview

AI is used in so many different contexts that only a flexible, risk management approach will be successful. The BSA Framework is built on three key elements:

(1) Identifying the risks of bias through **impact assessments** across a system's lifecycle;
(2) **Mitigating those risks** through concrete, actionable practices; and
(3) Setting forth key corporate governance structures to promote **organization accountability**.

Among the unique features of the BSA Framework is that it recognizes these elements need to be followed at all stages of the AI lifecycle: Design, Development, and Deployment and Use phases. Further, there are a variety of AI development and deployment models, and the Framework recognizes that the appropriate allocation of risk management responsibilities will vary depending on the type of system, including who develops the algorithm, trains the model, and ultimately deploys the system.

- ***AI Bias Can Arise Throughout the AI Lifecycle***

To combat AI bias, it is essential to understand the many sources of risk and the variety of ways they can manifest in an AI system. While much attention has understandably focused on data as a source of bias, the potential vectors of risk precede data collection efforts and begin at the earliest stages of a system's conception and design.

The initial step in building an AI system is often referred to as "problem formulation." It involves the identification and specification of the "problem" the system is intended to address, an initial mapping of how the model will achieve that objective, and the identification of a "target variable" the system will be used to predict. Because many AI systems are designed to make predictions about attributes that are not directly measurable, data scientists must often identify variables that can be used as proxies for the quality or outcome it is intended to predict.

While the use of proxy target variables can be entirely reasonable, the assumptions underlying the choice of proxies must be closely scrutinized to ensure that it does not introduce unintended bias to the system. The risk that can arise during this process of problem formulation is perhaps best exemplified by a recent study of a widely used healthcare algorithm that hospitals rely on to identify patients in need of urgent care. The research team concluded that the algorithm was systematically assigning lower risk scores to black patients compared to similarly sick white counterparts because it relied on data about historical healthcare costs as proxy for predicting a patient's future healthcare needs. Unfortunately, because black patients have historically had less access to healthcare, the reliance of spending data painted an inaccurate picture and led to dangerously biased outcomes.[6]

[6] Millions of black people affected by racial bias in health-care algorithms (nature.com)

The data used to train an AI system is a second major vector for bias. If the data used to train a system is misrepresentative of the population in which it will be used, there is a risk the system will perform less effectively on communities that may be underrepresented in the training data. Likewise, reliance on data that itself may be the product of institutional or historical biases can entrench those inequities in an AI model. The process of "labelling" training data can also introduce bias. Many AI systems require training data to be "labeled" so that the learning algorithm can identify patterns and correlations that can be used to classify future data inputs. Because the process of labeling the data can involve subjective decisions, there is the potential for introducing unintended bias into the training data.

Finally, even a system thoroughly vetted during development can begin to exhibit bias after it is deployed. AI systems are trained on data that represents a static moment in time and that filters out "noise" that could undermine the model's ability to make consistent and accurate predictions. Upon deployment in the real world, AI systems inevitably encounter conditions that differ from those in the development and testing environment. Further, because the real-world changes over time, the snapshot in time that a model represents may naturally become less accurate as the relationship between data variables evolves. If the input data for a deployed AI system differs materially from its training data, there is a risk that the system could "drift" and that the performance of the model could be undermined in ways that will exacerbate the risks of bias. For instance, if an AI system is designed (and tested) for use in a specific country, the system may not perform well if it is deployed in a country with radically different demographics. Bias can also arise if an AI system is deployed into an environment that differs significantly from the conditions for which it was designed or for purposes that are inconsistent with its intended use.

- ***Combatting AI Bias Requires a Lifecycle-Based Approach to Risk Management***

Although the challenges of AI bias are significant and without simple solutions, they are not insurmountable. Efforts to combat bias must start by recognizing that the issue requires a lifecycle-based approach to risk management.

Risk management is a process for ensuring systems are trustworthy by design by establishing a methodology for identifying risks and mitigating their potential impact. Risk management processes are particularly important in contexts, such as cybersecurity and privacy, where the combination of quickly evolving technologies and a highly dynamic threat landscapes render traditional "compliance" based approaches ineffective. Rather than evaluating a product or service against a static set of prescriptive requirements that quickly become outdated, risk management seeks to integrate compliance responsibilities into the development pipeline to help mitigate risks throughout a product or service's lifecycle.

But, what does that all mean in practice?

Companies that develop or use high-risk AI systems should establish a comprehensive approach for performing impact assessments. Impact assessments are widely used in a range of other fields—from environmental protection to data protection—as an accountability mechanism that promotes trust by demonstrating that a system has been designed in a manner that accounts for the potential risks it may pose to the public. The purpose of an impact assessment is to establish organizational processes to guide the development and use of high-risk systems by requiring internal stakeholders to identify the risks that a system may pose, quantify the degree of harm the system could generate, and document any steps that have been taken to mitigate those risks to an acceptable level. By establishing a process for personnel to document key design choices and their underlying rationale, impact assessments are an important transparency and accountability mechanism.

The impact assessment methodology in the BSA Framework includes more than 40 diagnostic statements that should be documented throughout an AI system's lifecycle. Among its key recommendations is for organizations to maintain documentation about:

- The objectives and assumptions of the system, including its intended use cases and its target variable;
- The metrics that will be used as a baseline for evaluating bias in the system;
- The provenance of the data used to train the system, an evaluation of its appropriateness for the intended use case, and the steps that were taken to scrutinize the data for biases;
- The rationale for selecting data attributes and their impact on model performance; and
- The lines of responsibility for monitoring the system following deployment and plans for responding to potential incidents or system errors.

- ***Risk Management is a Collective Responsibility***

The documentation created and maintained as part of an impact assessment also facilitates important communication between the multiple stakeholders that may have roles to play managing AI risks. In many instances, the risk of bias may emerge at the intersection of system design decisions that were made by the system's developer and downstream decisions by the organizations that may deploy that system.

For instance, some AI developers provide general-purpose AI functionality, such as text analytics tools, that their customers can access and integrate into their own products and services via an API. In such a circumstance, risk management responsibilities will necessarily be shared by the system developer and the organization that deployed it. In other situations, the customers may, for privacy or other purposes, not allow the developer to view or assess data that may be used to re-train or fine tune the AI model.

While the precise allocation of risk management responsibilities will vary depending on the use case, as a general matter AI developers will be best positioned to provide information about the system's design and capabilities to enable the deployer to make informed deployment and risk mitigation decisions.

- ***Mitigating AI Bias Requires Diverse, Interdisciplinary Expertise***

A common refrain in the BSA Framework relates to the vital role of diversity in AI risk management efforts. Effectively identifying potential sources of bias in data requires a diverse set of expertise and experiences, including familiarity with the domain from which data is drawn and a deep understanding of the historical context and institutions that produced it. Moreover, oversight processes are most effective when team members bring diverse perspectives and backgrounds that can help anticipate the needs and concerns of users who may be impacted by or interact with an AI system.

Because "algorithm development implicitly encodes developer assumptions that they may not be aware of, including ethical and political values," it is vital for organizations to establish teams that reflect a diversity of lived experiences and that traditionally underrepresented perspectives are included throughout the lifecycle of the AI design and development process.[7] To the extent an organization is lacking in diversity, it should consult with outside stakeholders to solicit feedback, particularly from underrepresented groups that may be impacted by the system.

Policy Recommendations

Public trust is an essential component of a thriving digital economy. While the responsibility for managing the risks of AI falls squarely on the organizations that develop and use AI systems, government can help foster public trust through policies that enhance the benefits of the technology while safeguarding against its potential risks. In the near term, we would advise Congress and the Administration to focus on the following lines of effort.

(1) Ensure consumer and civil rights protections remain fit-for-purpose in the digital age. Decisions that would otherwise be unlawful should not avoid liability simply because they may now involve the use of an AI system. To that end, we have encouraged efforts to audit federal agencies' existing consumer protection authorities to assess whether technological innovation is impeding their ability to enforce the law.[8] And we wrote to this Task Force last year about concerns that a rulemaking at HUD may exacerbate the risk of bias and discrimination in the housing market.[9]

(2) Establish a requirement for organizations to perform impact assessments prior to deploying high-risk AI systems. The BSA Framework can be one useful roadmap for new legislation.

(3) Promote international alignment around AI policy. Given the inherently global nature of the technology ecosystem, it is vital for the US to engage with our trading partners to forge consensus approaches for tackling shared challenges. There is an emerging global consensus that AI regulation should be risk-based and context specific. The EU recently introduced comprehensive legislation along these lines. The US should look for opportunities to drive these conversations, including through NIST's development of an AI risk management framework.

(4) Continue to emphasize privacy and security. Ethics and issues of bias are part of the trust formula, but privacy and data security laws are also essential.

[7] Inioluwa Deborah Raji et al., *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*, FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (January 2020): 33–44, https://doi.org/10.1145/3351095.3372873.

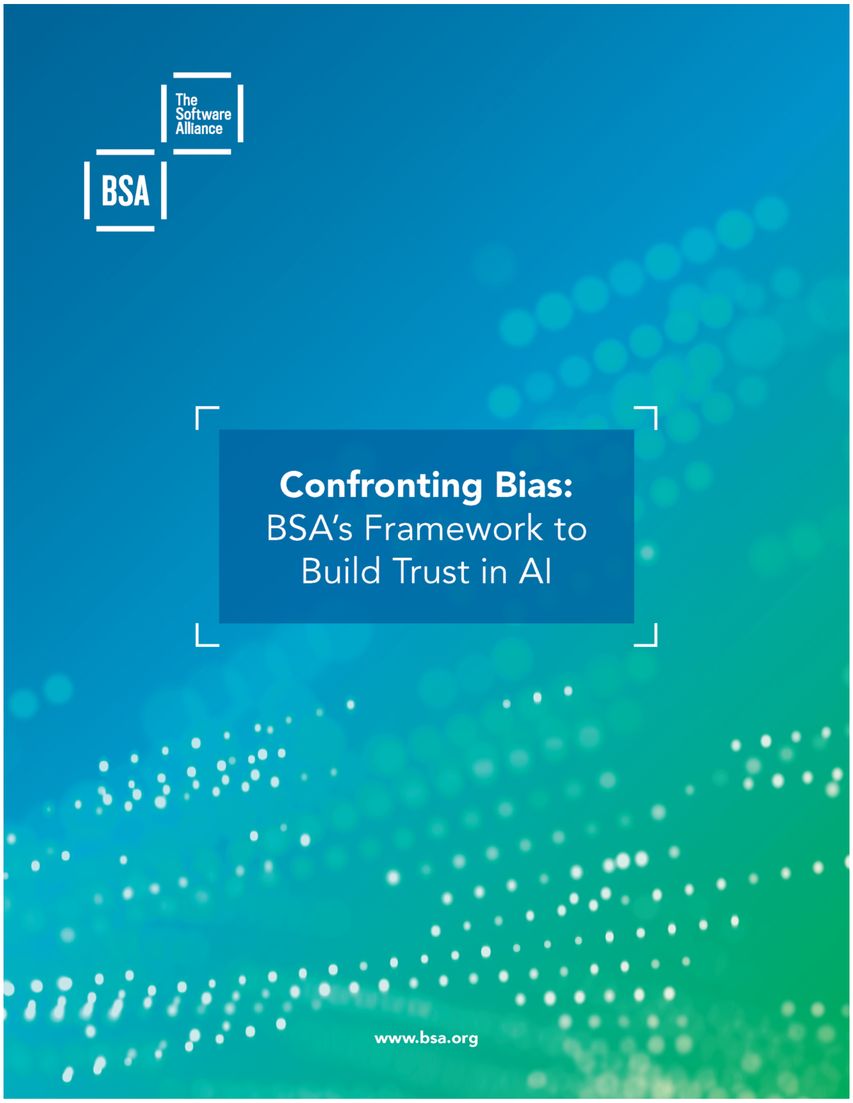
[8] *See* BSA Comments on Office of Management and Budget's Guidance on AI Regulation | BSA | The Software Alliance

[9] US: BSA Letter to the House Financial Services Committee Regarding Equitable Algorithms Hearing

Conclusion

Digital transformation across industry sectors is creating jobs and improving our lives. But industry, civil society, academia, and the government must work together on guidelines and laws that will ensure companies act responsibly in how they develop and deploy AI.

We appreciate the Task Force’s strong focus on issues of ethics and bias. Confronting Bias: BSA’s Framework to Build Trust in AI is our attempt to contribute meaningfully to this discussion. Thank you again for the opportunity to testify.



Confronting Bias:
BSA's Framework to
Build Trust in AI

www.bsa.org

CONTENTS

Introduction 1

What Is AI Bias? 3

Sources and Types of AI Bias 4

The Need for AI Risk Management 8

What Is Risk Management? 8

Managing the Risk of Bias 9

Foundations for Effective Risk Management 10

Governance Framework 11

Impact Assessment 13

AI Bias Risk Management Framework 14

AI Lifecycle Phases 15

Framework Structure 17

Stakeholder Roles and Responsibilities 18

Spectrum of AI Development and Deployment Models 18

BSA AI Bias Risk Management Framework 19

Foundational Resources 28

Endnotes 29

Introduction

Tremendous advances in artificial intelligence (AI) research and development are quickly transforming expectations about how the technology may shape the world. The promise that AI may one day impact every industry is quickly turning into a commercial reality. From financial services to healthcare, AI is increasingly leveraged to improve customer experiences, enhance competitiveness, and solve previously intractable problems. For instance, AI is enabling medical researchers to diagnose early-stage Alzheimer's Disease years before debilitating symptoms arise,[1] and it is helping ecologists analyze impossibly large datasets to better track the impact of their efforts to preserve critical habitat and prevent illegal elephant poaching in Malawi.[2]

As used in this report, the term "artificial intelligence" refers to systems that use machine learning algorithms that can analyze large volumes of training data to identify correlations, patterns, and other metadata that can be used to develop a model that can make predictions or recommendations based on future data inputs. For example, developers used machine learning to create "Seeing AI," an app that helps people who are blind or visually impaired navigate the world by providing auditory descriptions of objects in photographs.[3] Users of the app can use their smartphone to take pictures, and Seeing AI describes what appears in the photograph. To develop the computer vision model capable of identifying the objects in a picture, the system was trained using data from millions of publicly available images depicting common objects, such as trees, street signs, landscapes, and animals. When a user inputs a new image, Seeing AI in effect predicts what objects are in the photo by comparing it to the patterns and correlations that it derived from the training data.

The proliferation of AI across industries is also prompting questions about the design and use of the technology and what steps can be taken to ensure it is operating in a manner that accounts for any potential risks it may pose to the public.

The use of advanced technologies in connection with high-stakes decisions presents both opportunities and risks. On the one hand, the adoption of AI by financial institutions has the potential to reduce discrimination and promote fairness by facilitating a data-driven approach to decision-making that is less vulnerable to human biases.[4] For instance, the use of AI can improve access to credit and housing to historically marginalized communities by enabling lenders to evaluate a greater array of data than is ordinarily accounted for in traditional credit reports. At the same time, researchers caution that flaws in the design, development, and/or deployment of AI systems have the potential to perpetuate (or even exacerbate) existing societal biases.[5]

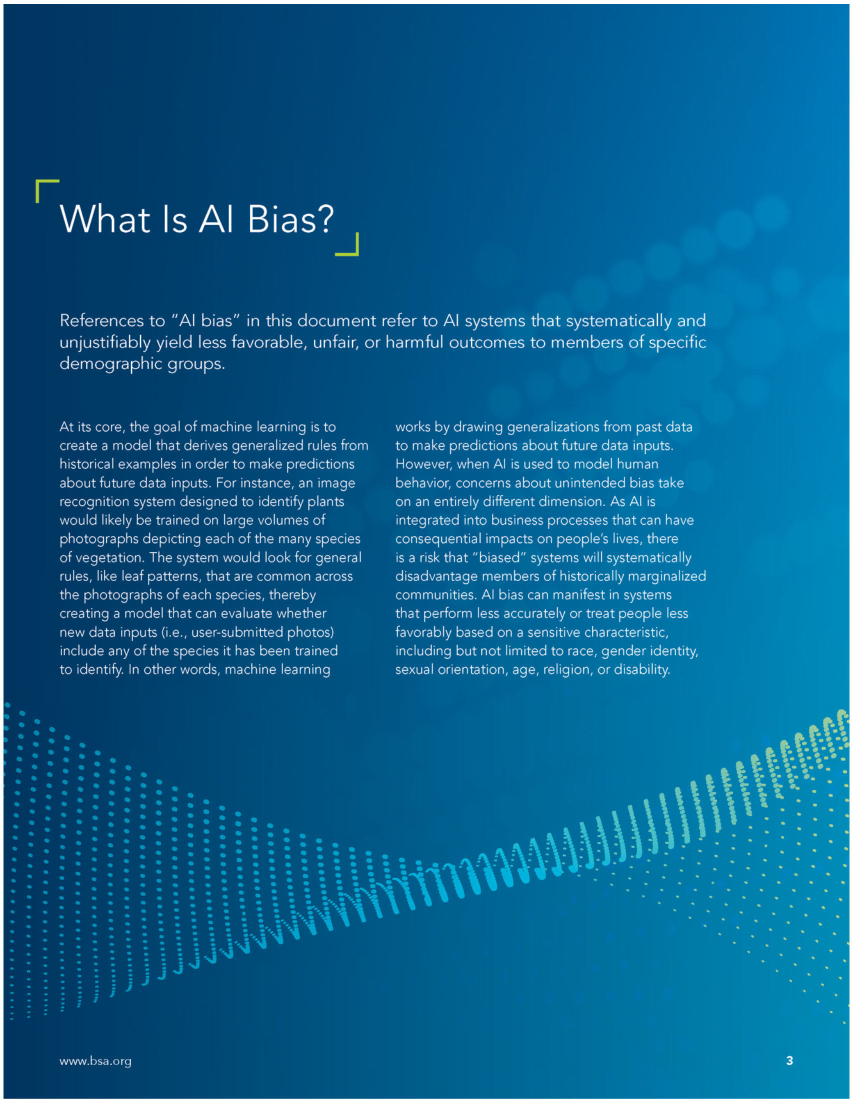
Developing mechanisms for identifying and mitigating the risks of AI bias has therefore emerged as an area of intense focus for experts in industry, academia, and government. In just the past few years, a vast body of research has identified a range of organizational best practices, governance safeguards, and technical tools that can help manage the risks of bias throughout the AI lifecycle. Static evaluations of AI models cannot account for all potential issues that may arise when AI systems are deployed in the field, so experts agree that mitigating risks of AI bias requires a lifecycle approach that includes ongoing monitoring by end-users to ensure that the system is operating as intended.

This document sets forth an *AI Bias Risk Management Framework* that organizations can use to perform impact assessments to identify and mitigate potential risks of bias that may emerge throughout an AI system's lifecycle. Similar to impact assessments for data privacy, AI impact assessments can serve as an important assurance mechanism that promotes accountability and enhances trust that high-risk AI systems have been designed, developed, tested, and deployed with sufficient protections in place to mitigate the risk of harm. AI impact assessments are also an important transparency mechanism that enables the many potential stakeholders involved in the design, development, and deployment of an AI system to communicate about its risks and ensure that responsibilities for mitigating those risks are clearly understood.

In addition to setting forth a process for performing an AI impact assessment, the Bias Risk Management Framework:

- Sets out the key corporate governance structures, processes, and safeguards that are needed to implement and support an effective AI risk management program; and
- Identifies existing best practices, technical tools, and resources that stakeholders can use to mitigate specific AI bias risks that can emerge throughout an AI system's lifecycle.

This Framework is intended to be a flexible tool that organizations can use to enhance trust in their AI systems through risk management processes that promote fairness, transparency, and accountability.



What Is AI Bias?

References to "AI bias" in this document refer to AI systems that systematically and unjustifiably yield less favorable, unfair, or harmful outcomes to members of specific demographic groups.

At its core, the goal of machine learning is to create a model that derives generalized rules from historical examples in order to make predictions about future data inputs. For instance, an image recognition system designed to identify plants would likely be trained on large volumes of photographs depicting each of the many species of vegetation. The system would look for general rules, like leaf patterns, that are common across the photographs of each species, thereby creating a model that can evaluate whether new data inputs (i.e., user-submitted photos) include any of the species it has been trained to identify. In other words, machine learning works by drawing generalizations from past data to make predictions about future data inputs. However, when AI is used to model human behavior, concerns about unintended bias take on an entirely different dimension. As AI is integrated into business processes that can have consequential impacts on people's lives, there is a risk that "biased" systems will systematically disadvantage members of historically marginalized communities. AI bias can manifest in systems that perform less accurately or treat people less favorably based on a sensitive characteristic, including but not limited to race, gender identity, sexual orientation, age, religion, or disability.

Sources and Types of AI Bias

DESIGN

AI bias can be introduced at multiple stages in the AI lifecycle.[6] Decisions made at the earliest stages of the conception and design of an AI system can introduce bias:

- **Problem Formulation Bias.** In some instances, the basic assumptions underlying a proposed AI system may be so inherently biased that they render it inappropriate for any form of public deployment.

EXAMPLES

In 2016, researchers at Shanghai Jiao Tong University published a highly controversial paper[7] detailing their effort to train an AI system to predict "criminality" through a facial imaging system. By training the system on a large volume of police mugshots, the researchers alleged that their system could predict "criminality" with close to 90 percent accuracy merely by analyzing a person's facial structure. Unsurprisingly, the paper quickly became the subject of scathing criticism, and commentators rightfully noted that the model relied on the profoundly disturbing (and causally unsupportable) assumption that criminality can be inferred from a person's appearance.[8]

Problem formulation bias can also arise when an AI system's target variable is an imprecise or overly simplistic proxy for what the system is actually trying to predict. For example, in 2019 researchers discovered that an AI system widely used by hospitals to triage patients[9] by predicting the likelihood that they required urgent care systematically prioritized the needs of healthier white patients to the detriment of less-healthy minority patients. In this instance, bias arose because the system sought to predict "healthcare needs" using historical data about "healthcare costs" as an easy-to-obtain stand-in for the actual data about the healthcare needs of patients. Unfortunately, because minority patients have historically had less access to healthcare, using "healthcare costs" as a proxy for the current needs of those patients paints an inaccurate picture that can result in dangerously biased outcomes.

- **Historical Bias.** There is a risk of perpetuating historical biases reflected in data used to train an AI system.

EXAMPLE

A medical school in the United Kingdom set out to create a system that would help identify good candidates for admission. The system was trained using data about previously admitted students. It was discovered, however, that the school's historical admissions decisions had systematically disfavored racial minorities and females whose credentials were otherwise equal to other applicants. By training the model using data reflecting historical biases, the medical school inadvertently created a system that replicated those same biased admission patterns.[10]

- **Sampling Bias.** If the data used to train a system is misrepresentative of the population in which it will be used, there is a risk that the system will perform less effectively on communities that may have been underrepresented in the training data. This commonly occurs when sufficient quantities of representative data are not readily available, or when data is selected or collected in ways that systematically over- or under-represent certain populations.

EXAMPLES

As the pathbreaking research by Joy Buolamwini and Timnit Gebru demonstrated, facial recognition systems trained on datasets composed disproportionately of white and male faces perform substantially less accurately when evaluating the faces of women with darker complexions.[11]

Sampling bias can also arise as a result of data collection practices. The City of Boston's attempt to create a system capable of automatically detecting and reporting potholes in need of repair is an illustrative case in point. Because early versions of the program relied heavily on data supplied by users of a smartphone app called "StreetBump," it received a disproportionate number of reports from affluent neighborhoods with residents who could afford smartphones and data plans. As a result of the sampling bias, potholes in poorer neighborhoods were underrepresented in the dataset, creating a risk that the system would allocate repair resources in a manner that would treat members of those communities unfairly.[12]

- **Labeling Bias.** Many AI systems require training data to be "labeled" so that the learning algorithm can identify patterns and correlations that can be used to classify future data inputs. The process of labeling the training dataset can involve subjective decisions that can be a vector for introducing human biases into the AI system.

EXAMPLE

ImageNet is a database of more than 14 million images that have been categorized and labeled to enable AI researchers to train vision recognition systems. Although ImageNet has been a critical tool for advancing the state of the art in AI object recognition, recent scholarship has shone a light on how the database's categorization and labeling system can create significant risks of bias when it is used to train systems involving images of people. In *Excavating AI*,[13] Kate Crawford and Trevor Paglen demonstrated that the categories and data labels associated with the images of people in ImageNet reflect a range of "gendered, racialized, ableist, and ageist" biases that could be propagated in any AI system that uses them as training data. For instance, an AI system trained on ImageNet data was more likely to classify images of Black subjects as "wrongdoers" or "offenders."[14]

DEVELOPMENT

Once the necessary data has been collected, the development team must clean, process, and normalize the data so that it can be used to train and validate a model. Developers must also select a machine learning approach, or adapt an off-the-shelf model, that is appropriate for the nature of the data they are using and the problem they are trying to solve. This may involve building many different models using different approaches and then choosing the most successful among them.[15] Usually, the development team must also make choices about data parameters to make the model functional. For instance, data reflecting a numerical score may be converted to a "yes" or "no" answer by assigning a threshold—for example, scores equal or greater to X may be re-designated as a "yes," and scores below that threshold designated "no." Biases that can emerge during the development stage include the following:

- **Proxy Bias.** The process of selecting the input variables (i.e., "features") that the model will weigh as it is being trained is another critical decision point that can introduce bias. Even when sensitive demographic data is excluded, bias may be introduced if the system relies on features that are closely correlated to those traits, called proxies.

EXAMPLE

Even the use of seemingly benign features can introduce proxy bias due to their correlation with sensitive attributes. Researchers have shown, for instance, that information about whether a person owns a Mac or PC laptop may be predictive of their likelihood to pay back a loan.[16] A financial institution might therefore seek to include such a variable when building an AI system to screen potential loan applicants. However, the inclusion of that feature also introduces a significant risk of proxy bias because Mac ownership correlates closely to race. As a result, its inclusion could result in a system that systematically disfavors applicants based on a feature that is closely correlated to race but that is unrelated to actual credit risk.

- **Aggregation Bias.** Using a "one-size-fits-all" model that overlooks key variables can result in system performance that is optimized only for the dominant sub-group. Aggregation bias can arise if the model fails to account for underlying differences between sub-groups that materially impact a system's accuracy rates. Rare phenomena may be lost in averages and aggregates. Worse, models of aggregated populations may correctly predict different or even opposite behavior to modes of sub-groups of the same population, a phenomenon known as Simpson's Paradox.

EXAMPLE

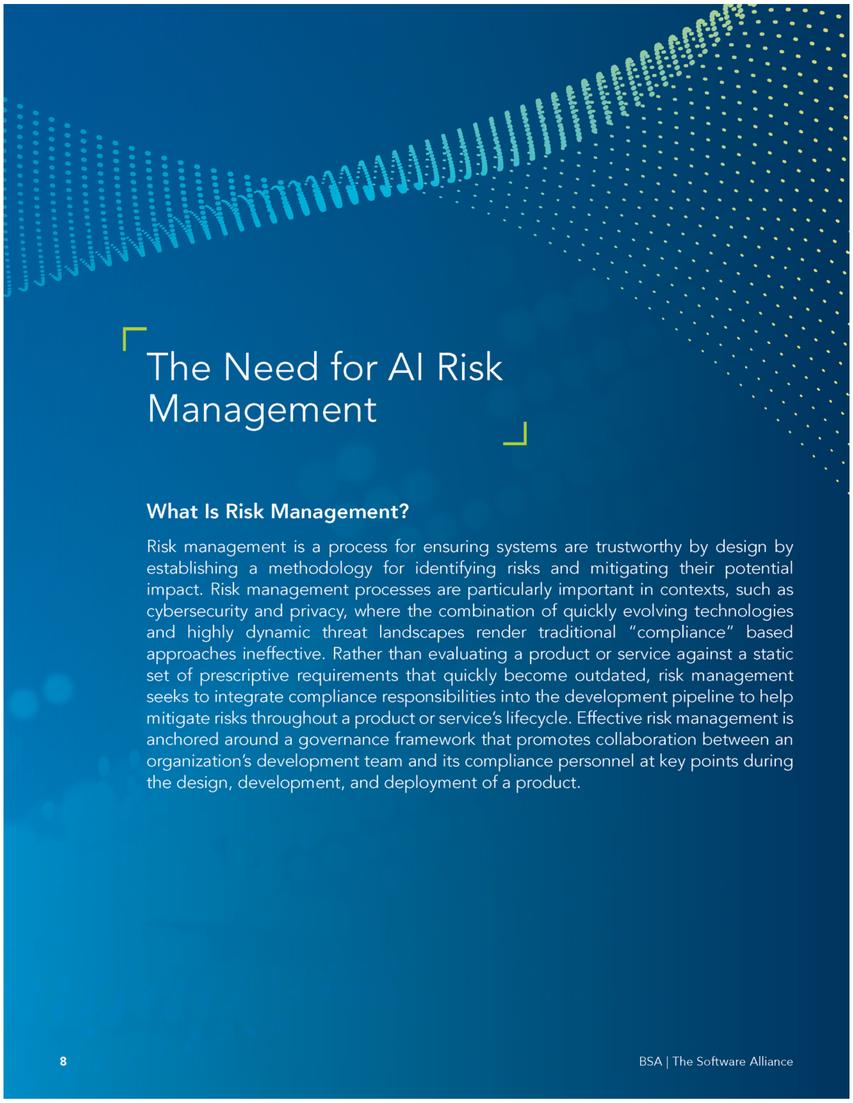
The risk of aggregation bias is particularly acute in healthcare settings where diagnosis and treatment must often account for the unique manner in which medical conditions may impact people across racial and ethnic lines. For instance, because the risk of complications posed by diabetes varies wildly across ethnicities, an AI system used to predict the risks associated with diabetes may underperform for certain patients unless it accounts for these differences.[17]

DEPLOYMENT, MONITORING, AND ITERATION

AI systems inevitably encounter real world scenarios that differ from the data used to train the model. As a result, even a system that has been thoroughly validated and tested prior to deployment may suffer performance degradation when it is put into production. Therefore, it is important that AI systems undergo ongoing evaluation and assessment throughout their lifecycles.

- **Deployment Bias.** Bias can arise in various ways after a system has been deployed, including when the data used to train or evaluate an AI system differs markedly from the population the system encounters when it is deployed, rendering the model unable to perform as intended. Deployment bias can emerge when a model is unable to reliably generalize beyond the data on which it was trained, either because the model was overfitted at the time of training (i.e., the prediction model learned so much detail about the training data that it is unable to make accurate generalizations about other data inputs) or because of concept drift (i.e., performance degradation was brought on by a shift in the relationship between the target variable and the training data).

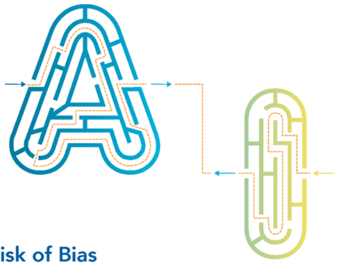
- **Misuse Bias.** Deployment bias can also arise when an AI system or feature built for one purpose is used in an unexpected or unintended manner.



The Need for AI Risk Management

What Is Risk Management?

Risk management is a process for ensuring systems are trustworthy by design by establishing a methodology for identifying risks and mitigating their potential impact. Risk management processes are particularly important in contexts, such as cybersecurity and privacy, where the combination of quickly evolving technologies and highly dynamic threat landscapes render traditional "compliance" based approaches ineffective. Rather than evaluating a product or service against a static set of prescriptive requirements that quickly become outdated, risk management seeks to integrate compliance responsibilities into the development pipeline to help mitigate risks throughout a product or service's lifecycle. Effective risk management is anchored around a governance framework that promotes collaboration between an organization's development team and its compliance personnel at key points during the design, development, and deployment of a product.



Managing the Risk of Bias

Organizations that develop and use AI systems must take steps to prevent bias from manifesting in a manner that unjustifiably yields less favorable or harmful outcomes based on someone's demographic characteristics. Effectively guarding against the harms that might arise from such bias requires a risk management approach because:

"BIAS" AND "FAIRNESS" ARE CONTEXTUAL

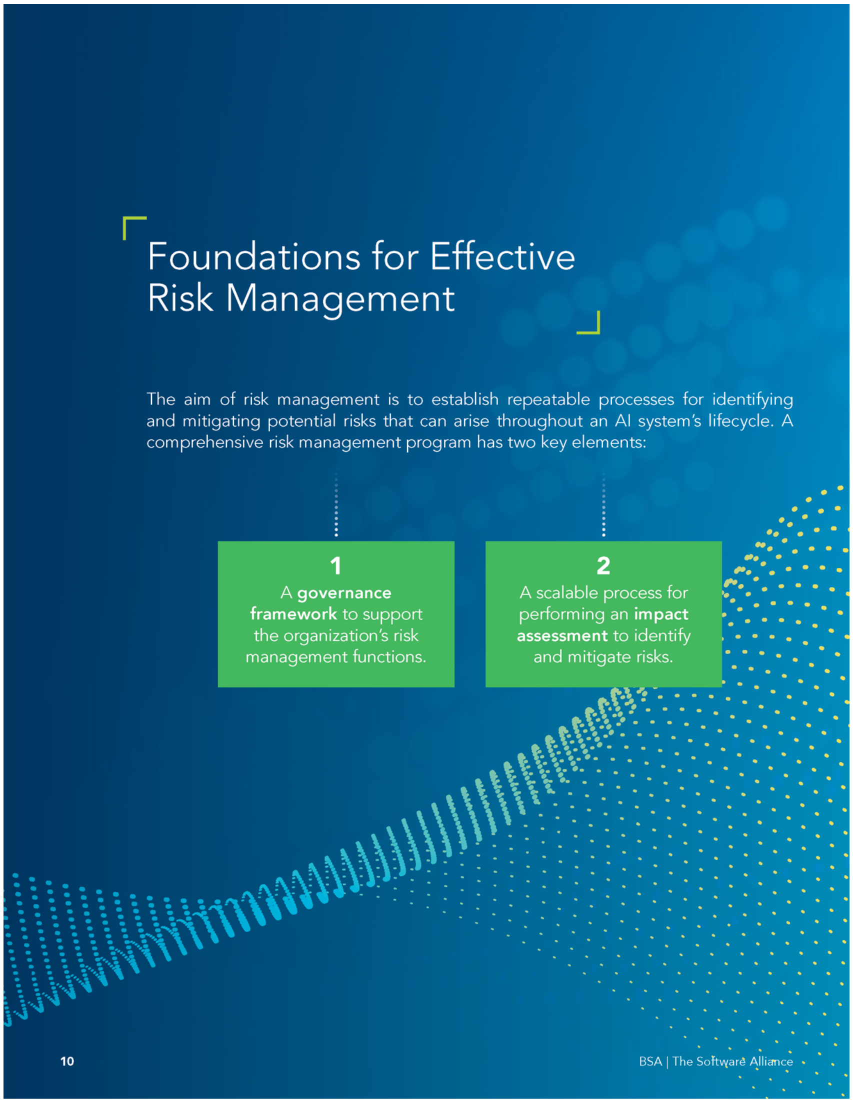
It is impossible to eliminate bias from AI systems because there is no universally agreed upon method for evaluating whether a system is operating in a manner that is "fair." In fact, as Professor Arvind Narayanan has famously explained, there are at least 21 different definitions[18] (i.e., mathematical criteria) that can be used to evaluate whether a system is operating fairly, and it is *impossible* for an AI system to simultaneously satisfy all of them. Because no universal definition of fairness exists, developers must instead evaluate the nature of the system they are creating to determine which metric for evaluating bias is most appropriate for mitigating the risks that it might pose.

EFFORTS TO MITIGATE BIAS MAY INVOLVE TRADE-OFFS

Interventions to mitigate bias for one group can increase it for other groups and/or reduce a system's overall accuracy.[19] Risk management provides a mechanism for navigating such trade-offs in a context-appropriate manner.

BIAS CAN ARISE POST-DEPLOYMENT

Even if a system has been thoroughly evaluated prior to deployment, it may produce biased results if it is misused or deployed in a setting in which the demographic distribution differs from the composition of its training and testing data.



Foundations for Effective Risk Management

The aim of risk management is to establish repeatable processes for identifying and mitigating potential risks that can arise throughout an AI system's lifecycle. A comprehensive risk management program has two key elements:

1

A **governance framework** to support the organization's risk management functions.

2

A scalable process for performing an **impact assessment** to identify and mitigate risks.

Governance Framework

Effective AI risk management should be underpinned by a governance framework that establishes the policies, processes, and personnel that will be used to identify, mitigate, and document risks throughout the system's lifecycle. The purpose of such a governance framework is to promote understanding across organizational units—including product development, compliance, marketing, sales, and senior management—about each entity's role and responsibilities for promoting effective risk management during the design, development, and deployment of AI systems. Key features of a risk management governance framework include:

Policies and Processes

At the core of the governance framework is a set of formal policies setting forth the organization's approach to risk management. These policies should define the organization's risk management objectives, the procedures that it will use to meet those objectives, and the benchmarks it will rely on for evaluating compliance.

- **Objectives.** AI risk management should be contextualized within an organization's broader risk management functions with the goal of ensuring that the organization is developing and using AI in a manner that aligns with its core values. To that end, the governance framework should identify how the organization will manage risks that could undermine those values.
- **Processes.** The governance framework should establish processes and procedures for identifying risks, assessing the materiality of those risks, and mitigating risks at each stage of the AI lifecycle.
- **Evaluation Mechanisms.** The governance framework should establish mechanisms, such as metrics and benchmarks, that the organization will use to evaluate whether policies and procedures are being carried out as specified.
- **Periodic Review.** As AI capabilities continue to mature and the technology is put to new uses, it is important that organizations periodically review and update their AI governance framework so that it remains fit-for-purpose and capable of addressing the evolving landscape of risk.

Executive Oversight. AI Developers and AI Deployers should maintain a governance framework that is backed by sufficient executive oversight. In addition to developing and approving the substance of the governance framework's policies, senior management should play an active role in overseeing the company's AI product development lifecycle. For high-risk systems that may negatively impact people in consequential ways, company leadership should be accountable for making "go/no-go" decisions.

Personnel, Roles, and Responsibilities

The effectiveness of risk management depends on establishing a cross-functional group of experts that can guide decisions throughout the AI lifecycle. Depending on the size of an organization and the nature of the systems it is developing or deploying, the responsibilities for risk management may involve staff from multiple business units. The governance framework should therefore identify the personnel within the organization who have roles and responsibilities related to AI risk management and clearly map reporting lines, authorities, and necessary expertise. In assigning roles and responsibilities, organizations should prioritize independence, competence, influence, and diversity.

- **Independence.** Risk management is most effective when personnel are structured in a manner that facilitates separate layers of independent review. For instance, risk management responsibilities may be split between multiple teams, including:
  - **Product Development Team.** Engineers, data scientists, and domain experts involved in designing and developing AI products and services.
  - **Compliance Team.** A diverse team of legal, compliance, domain experts, and data professionals who are responsible for overseeing compliance with the company's AI development policies and practices, such as the development of impact assessments for high-risk AI systems.
  - **Governance Team.** Ideally a senior management-led team with responsibility for developing, maintaining, and ensuring effective oversight of the organization's AI Governance Framework and risk management processes.
- **Competence, Resourcing, and Influence.** Personnel with risk management responsibilities must be provided with adequate training and resources to fulfill their governance functions. It is equally important to ensure that personnel are empowered and have the right incentives to make decisions to address and/or escalate risks. For instance, the organization should establish a clear escalation path that enables risk management personnel to engage with executive decision-makers so that there is executive-level visibility into key risk areas and decisions.

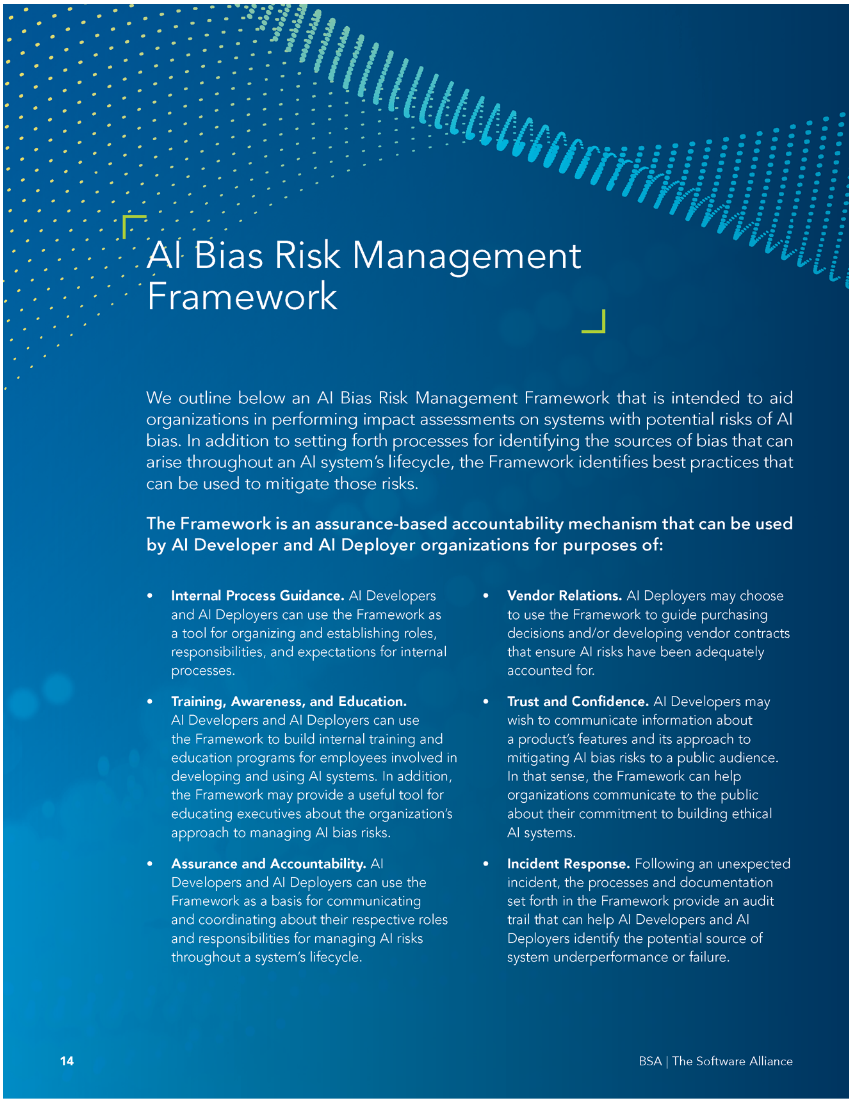
Diversity. The sociotechnical nature of AI systems makes it vitally important to prioritize diversity within the teams involved in a system's development and oversight. Development and oversight processes are most effective when team members bring diverse perspectives and backgrounds that can help anticipate the needs and concerns of users who may be impacted by or interact with an AI system. Because "algorithm development implicitly encodes developer assumptions that they may not be aware of, including ethical and political values," it is vital that organizations establish teams that reflect a diversity of lived experiences and that traditionally underrepresented perspectives are included throughout the lifecycle of the AI design and development process.[20] To the extent an organization is lacking in diversity, it should consult with outside stakeholders to solicit feedback, particularly from underrepresented groups that may be impacted by the system.

Impact Assessment

To effectively manage AI risks, organizations should implement a robust process for performing impact assessments on any system that may materially impact members of the public. Impact assessments are widely used in a range of other fields—from environmental protection to data protection—as an accountability mechanism that promotes trust by demonstrating that a system has been designed in a manner that accounts for the potential risks it may pose to the public. In short, the purpose of an impact assessment is to identify the risks that a system may pose, quantify the degree of harm the system could generate, and document any steps that have been taken to mitigate those risks to an acceptable level.

Impact assessment processes should be tailored to address the nature of the system that is being evaluated and the type of harms it may pose. For truly low-risk systems—for example, a system used to predict the type of fonts being used on a document—a full impact assessment may not be necessary. But for systems that pose an inherent risk of material harm to the public, a full impact assessment should be performed. Given the incredible range of applications to which AI can be applied, there is no "one-size-fits-all" approach for identifying and mitigating risks. Instead, impact assessment processes should be tailored to address the nature of an AI system and the type of inherent risks and potential harms it may pose. To determine whether a system poses an inherent risk of material harm, stakeholders should consider:

- **Potential Impact on People.** Impact assessments are likewise important in circumstances where an AI system will be used in decision-making processes that may result in consequential impacts on people, such as their ability to obtain access to credit or housing.
- **Context and Purpose of the System.** Evaluating the nature of the AI system and the setting in which it will be used is a good starting point for determining both the necessity and appropriate scope of an impact assessment. Impact assessments are particularly critical for high-risk AI systems that will be used in domains (e.g., healthcare, transportation, finance) where the severity and/or likelihood of potential harms is high.
- **Degree of Human Oversight.** The degree to which an AI system is fully automated may also impact the inherent risks that it poses. A system designed to provide recommendations to a highly skilled professional is likely to pose fewer inherent risks than a similarly situated fully automated system. Of course, the mere existence of a human-in-the-loop certainly does not mean that an AI system is free from risk. It is necessary instead to examine the nature of the human-computer interaction holistically to determine the extent to which human oversight may mitigate an AI system's inherent risks.
- **Type of Data.** The nature of the data used to train a system can also shed light on a system's inherent risks. For instance, using training data relating to human characteristics or behaviors is a signal that a system may require closer scrutiny for bias.



AI Bias Risk Management Framework

We outline below an AI Bias Risk Management Framework that is intended to aid organizations in performing impact assessments on systems with potential risks of AI bias. In addition to setting forth processes for identifying the sources of bias that can arise throughout an AI system's lifecycle, the Framework identifies best practices that can be used to mitigate those risks.

The Framework is an assurance-based accountability mechanism that can be used by AI Developer and AI Deployer organizations for purposes of:

- **Internal Process Guidance.** AI Developers and AI Deployers can use the Framework as a tool for organizing and establishing roles, responsibilities, and expectations for internal processes.
- **Training, Awareness, and Education.** AI Developers and AI Deployers can use the Framework to build internal training and education programs for employees involved in developing and using AI systems. In addition, the Framework may provide a useful tool for educating executives about the organization's approach to managing AI bias risks.
- **Assurance and Accountability.** AI Developers and AI Deployers can use the Framework as a basis for communicating and coordinating about their respective roles and responsibilities for managing AI risks throughout a system's lifecycle.
- **Vendor Relations.** AI Deployers may choose to use the Framework to guide purchasing decisions and/or developing vendor contracts that ensure AI risks have been adequately accounted for.
- **Trust and Confidence.** AI Developers may wish to communicate information about a product's features and its approach to mitigating AI bias risks to a public audience. In that sense, the Framework can help organizations communicate to the public about their commitment to building ethical AI systems.
- **Incident Response.** Following an unexpected incident, the processes and documentation set forth in the Framework provide an audit trail that can help AI Developers and AI Deployers identify the potential source of system underperformance or failure.

AI Lifecycle Phases

The Framework is organized around the phases of the AI lifecycle, which represent the key iterative steps involved in the creation and use of an AI system.

DESIGN PHASE

- **Project Conception.** The initial stage of AI design involves identifying and formulating the "problem" that the system is intended to address and initially mapping how the model will achieve that objective. During this phase, the design team will define the purpose and structure of the system. Depending on the nature of the system, the design team will identify a target variable that the system is intended to predict. For instance, a fitness app that analyzes a consumer's heart rate to monitor for irregularities that might predict whether that person is at risk of a stroke or heart disease (i.e., the target variable). At this early stage of the system design process, the goal of the Bias Risk Management Framework is to identify whether using AI is appropriate for the project at hand. Potential risks include:
    - **Problem Formulation Bias.** Target variables may reflect inherent prejudices or faulty assumptions that can perpetuate harmful biases. In some instances, the basic assumptions underlying a proposed AI system may be so inherently biased as to render it inappropriate for any form of public deployment.
- **Data Acquisition.** Once the system objectives have been defined, developers must assemble a corpus of data that will be used to train the model to identify patterns that will enable it to make predictions about future data inputs. This training data can inadvertently introduce biases into an AI system in many ways. Potential risks include:
    - **Historical Bias.** Training an AI system using data that itself may reflect historical biases creates a risk of further entrenching those inequities.
    - **Sampling Bias.** The risk of bias also arises when the data used to train an AI system is not representative of the population in which it will be deployed. An AI system trained on unrepresentative data may not operate as effectively when making predictions about a member of a class that is either over- or under-represented.
    - **Labeling Bias.** Many AI systems require training data to be labeled so that it can identify what patterns it should be looking for. The process of labeling the training dataset can be a vector for introducing bias into the AI system.

DEVELOPMENT PHASE

- **Data Preparation and Model Definition.** The next step of the AI lifecycle involves preparing the data so that it is ready to train the model. During this process, the development team will clean, normalize, and identify the variables (i.e., "features") in the training data that the algorithm will evaluate as it looks for patterns and relationships as the basis of a rule for making future predictions. The team must also establish the system's underlying architecture, including selecting the type of algorithmic model that will power the system (e.g., linear regression, logistic regression, deep neural network.)[21] Once the data is ready and the algorithm is selected, the team will train the system to produce a functional model that can make predictions about future data inputs. Potential risks include the following:
  - **Proxy Bias.** The process of selecting features in the training data and choosing a modeling approach involves human decisions about what variables should be considered as relevant for making predictions about the model's target variable. These interventions can inadvertently introduce bias to the system, including by relying on variables that act as proxies for protected classes.
  - **Aggregation Bias.** Aggregation bias can arise if the model fails to account for underlying differences between sub-groups that materially impact a system's accuracy rates. Using a "one-size-fits-all" model that overlooks key variables can result in system performance that is optimized only for the dominant sub-group.
- **Model Validation, Testing, and Revision.** After the model has been trained, it must be validated to determine if it is operating as intended and tested to demonstrate that the system's outputs do not reflect unintended bias. Based on outcome of validation and testing, the model may need to be revised to mitigate risks of bias that are deemed unacceptable.

DEPLOYMENT PHASE

- **Deployment and Use.** Prior to deployment, the AI Developer should evaluate the system to determine whether risks identified in earlier stages of design and development have been sufficiently mitigated in a manner that corresponds to the company's governance policies. To the extent identified risks may arise through misuse of the system, the AI Developer should seek to control for them by integrating product features (e.g., user interfaces that reduce risk of misuse) to mitigate those risks, prohibiting uses that could exacerbate risks (e.g., end-user license agreements), and providing AI Deployers with sufficient documentation to perform their own impact assessments.

  Prior to using an AI system, an AI Deployer should review documentation provided by the AI Developer to assess whether the system corresponds with its own AI governance policies and to determine whether deployment-related risk management responsibilities are clearly assigned.

Although some post-deployment risk management responsibilities may be addressed by the AI Developer, the AI Deployer will often bear responsibility for monitoring system performance and evaluating whether it is operating in a manner that is consistent with its risk profile. Potential risks include:

- **Deployment Bias.** AI systems are trained on data that represents a static moment in time and that filters out "noise" that could undermine the model's ability to make consistent and accurate predictions. Upon deployment in the real world, AI systems will necessarily encounter conditions that differ from those in the development and testing environment. Further, because the real-world changes over time, the snapshot in time that a model represents may naturally become less accurate as the relationship between data variables evolves. If the input data for a deployed AI system differs materially from its training data, there is a risk that the system could "drift" and that the performance of the model could be undermined in ways that will exacerbate the risks of bias. For instance, if an AI system is designed (and tested) for use in a specific country, the system may not perform well if it is deployed in a country with radically different demographics.
- **Misuse Bias.** Deploying an AI system into an environment that differs significantly from the conditions for which it was designed or for purposes that are inconsistent with its intended use cases can exacerbate risks of bias.

Framework Structure

The Framework identifies best practices for identifying and mitigating risks of AI bias across the entire system lifecycle. It is organized into:

- **Functions,** which denote fundamental AI risk management activities at their highest level, dividing them between Impact Assessment and Risk Mitigation Best Practices.
- **Categories,** which set out the activities and processes that are needed to execute upon the Functions at each phase of the AI Lifecycle. In other words, the Categories set forth the steps for performing an Impact Assessment and identify the corresponding Risk Mitigation Best Practices that can be used to manage associated risks.
- **Diagnostic Statements,** which set forth the discrete actions that should be taken to execute upon the Categories. They provide a set of results that help support achievement of the outcomes in each Category.
- **Comments on Implementation,** which provide additional information for achieving the outcomes described in the Diagnostic Statements.
- **Tools and Resources,** which identify a range of external guidance and toolkits that stakeholders can use to mitigate the bias risks associated with each phase of the AI lifecycle. The specific tools and resources identified in the framework are non-exhaustive and are highlighted for informational purposes only.

Stakeholder Roles and Responsibilities

Reflecting the inherently dynamic nature of AI systems, the Framework is intended to account for the array of stakeholders that may play a role in various aspects of a system's design, development, and deployment. Because there is no single model of AI development or deployment, it is impossible in the abstract to assign roles or delegate specific responsibilities for many of the Framework's risk management functions. However, in general, there are three sets of stakeholders that may bear varying degrees of responsibility for certain aspects of AI risk management throughout a system's lifecycle:

- **AI Developers.** AI Developers are organizations responsible for the design and development of AI systems.
- **AI Deployers.** AI Deployers are the organizations that adopt and use AI systems. (If an entity develops its own system, it is both the AI Developer and the AI Deployer.)
- **AI End-Users.** AI End-Users are the individuals—oftentimes an employee of an AI Deployer—who are responsible for overseeing the use of an AI system.

The allocation of risk management responsibilities between these stakeholders will in many cases depend on an AI system's development and deployment model.

Spectrum of AI Development and Deployment Models

The appropriate allocation of risk management responsibilities between stakeholders will vary depending on the nature of the AI system being developed and which party determines the purposes and means by which the underlying model is trained. For instance:

- **Universal, Static Model.** The AI Developer provides all its customers (i.e., AI Deployers) with a static, pre-trained model.
  - The AI Developer will bear responsibility for most aspects of model risk management.
- **Customizable Model.** The AI Developer provides a pre-trained model to AI Deployers who can customize and/or retrain the model using their own data.
  - Risk management will be a shared responsibility between the AI Developer and the AI Deployer.
- **Bespoke Model.** The AI Developer trains a bespoke AI model on behalf of an AI Deployer using the AI Deployer's data.
  - Risk management will be a shared responsibility between the AI Developer and the AI Deployer, with the bulk of obligations falling on the AI Deployer.

BSA AI Bias Risk Management Framework

DESIGN

Function	Category	Diagnostic Statement	Comments on Implementation
PROJECT CONCEPTION			
Impact Assessment	Identify and Document Objectives and Assumptions	Document the intent and purpose of the system.	• What is the purpose of the system—i.e., what "problem" will it solve? • Who is the intended user of the system? • Where and how will the system be used? • What are the potential misuses?
		Clearly define the model's intended effects.	What is the model intended to predict, classify, recommend, rank, or discover?
		Clearly define intended use cases and context in which the system will be deployed.	
	Select and Document Metrics for Evaluating Fairness	Identify "fairness" metrics that will be used as a baseline for assessing bias in the AI system.	The concept of "fairness" is highly subjective and there are dozens of metrics by which it can be evaluated. Because it is impossible to simultaneously satisfy all fairness metrics, it is necessary to select metrics that are most appropriate for the nature of the AI system that is being developed and consistent with any applicable legal requirements. It is important to document the rationale by which fairness metrics were selected and/or excluded to inform latter stages of the AI lifecycle.
	Document Stakeholder Impacts	Identify stakeholder groups that may be impacted by the system.	Stakeholder groups include AI Deployers, AI End-Users, Affected Individuals (i.e., members of the public who may interact with or be impacted by an AI system).
		For each stakeholder group, document the potential benefits and potential adverse impacts, considering both the intended uses and reasonably foreseeable misuses of the system.	
		Assess whether the nature of the system makes it prone to potential bias-related harms based on user demographics.	User demographics may include, but are not limited to race, gender, age, disability status, and their intersections.
	Document Risk Mitigations	If risk of bias is present, document efforts to mitigate risks.	

DESIGN

Function	Category	Diagnostic Statement	Comments on Implementation
PROJECT CONCEPTION			
Impact Assessment *(continued)*	Document Risk Mitigations	Document how identified risks and potential harms of each risk will be measured and how the effectiveness of mitigation strategies will be evaluated.	
		If risk of bias is present, document efforts to mitigate risks.	
		If risks are unmitigated, document why the risk was deemed acceptable.	
Risk Mitigation Best Practices	Independence and Diversity	Seek feedback from a diverse set of stakeholders to inform the impact assessment.	Because risks identified during this initial phase will inform later aspects of the development and impact assessment processes, it is vital to develop a holistic understanding of potential harms that may arise by soliciting diverse perspectives from people with a range of lived experiences, cultural backgrounds, and subject matter expertise. To the extent in-house personnel lack subject matter or cultural diversity, it may be necessary to consult with third-party experts or to solicit feedback from members of communities that may be adversely impacted by the system.
	Transparent Documentation	Share impact assessment documentation with personnel working on later stages of the AI pipeline so that risks and potential unintended impacts can be monitored throughout the development process.	
	Accountability and Governance	Ensure that senior leadership has been adquately briefed on potential high risk AI systems.	Impact assessment documentation for systems deemed "high risk" should be shared with senior leadership to facilitate a "go/no-go" decision.
DATA ACQUISITION			
Impact Assessment	Maintain Records of Data Provenance	Maintain sufficient records to enable "recreation" of the data used to train the AI model, verify that its results are reproducible, and monitor for material updates to data sources.	Records should include: • Source of data • Origin of data (e.g., Who created it? When? For what purpose? How was it created?) • Intended uses and/or restrictions of the data and data governance rules (e.g., What entity owns the data? How long can it be retained (or must it be destroyed)? Are there restrictions on its use?) • Known limitations of data (e.g., missing elements?) • If data is sampled, what was the sampling strategy? • Will the data be updated? If so, will any versions be tracked?

DESIGN

Function	Category	Diagnostic Statement	Comments on Implementation
DATA ACQUISITION			
Impact Assessment *(continued)*	Examine Data for Potential Biases	Scrutinize data for historical biases.	Examine sources of data and assess potential that they may reflect historical biases.
		Evaluate "representativeness" of the data.	• Compare demographic distribution of training data to the population where the system will be deployed. • Assess whether there is sufficient representation of subpopulations that are likely to interact with the system.
		Scrutinize data labeling methodology.	• Document personnel and processes used to label data. • For third-party data, scrutinize labeling (and associated methodologies) for potential sources of bias.
	Document Risk Mitigations	Document whether and how data was augmented, manipulated, or re-balanced to mitigate bias.	
Risk Mitigation Best Practices	Independence and Diversity	To facilitate robust interrogation of the datasets, data review teams should include personnel that are diverse in terms of their subject matter expertise and lived experiences.	Effectively identifying potential sources of bias in data requires a diverse set of expertise and experiences, including familiarity with the domain from which data is drawn and a deep understanding of the historical context and institutions that produced it. To the extent in-house personnel lack diversity, consultation with third-party experts or potentially affected stakeholder groups may be necessary.
	Re-Balancing Unrepresentative Data	Consider re-balancing with additional data.	Improving representativeness can be achieved in some circumstances by collecting additional data that improves the balance of the overall training dataset.
		Consider re-balancing with synthetic data.	Imbalanced datasets can potentially be rebalanced by "oversampling" data from the underrepresented groups. A common oversampling method is the Synthetic Minority Oversampling Technique, which generates new "synthesized" data from the underrepresented group.

DESIGN

Function	Category	Diagnostic Statement	Comments on Implementation
DATA ACQUISITION			
Risk Mitigation Best Practices *(continued)*	Data Labeling	Establish objective and scalable labeling guidelines.	• To mitigate the potential of labeling bias, the personnel responsible for labeling the data should be provided with clear guidelines establishing an objective and repeatable process for individual labeling decisions. • In domains where the risk of bias is high, labelers should have adequate subject matter expertise and be provided training to recognize potential unconscious biases. • For high-risk systems, it may be necessary to set up a quality assurance mechanism to monitor label quality.
	Accountability and Governance	Integrate data labeling processes into a comprehensive data strategy.	Establishing an organizational data strategy can help ensure that data evaluation is performed consistently and prevent duplication of effort by ensuring that company efforts to scrutinize data are documented for future reference.

DESIGN: RISK MITIGATION TOOLS AND RESOURCES

Project Conception

- ***Aequitas Bias and Fairness Audit Toolkit***
  Pedro Saleiro, Abby Stevens, Ari Anisfeld, and Rayid Ghani, University of Chicago Center for Data Science and Public Policy (2018), http://www.datasciencepublicpolicy.org/projects/aequitas/.
- ***Diverse Voices Project | A How-To Guide for Facilitating Inclusiveness in Tech Policy***
  Lassana Magassa, Meg Young, and Batya Friedman, University of Washington Tech Policy Lab, https://techpolicylab.uw.edu/project/diverse-voices/.

Data Compilation

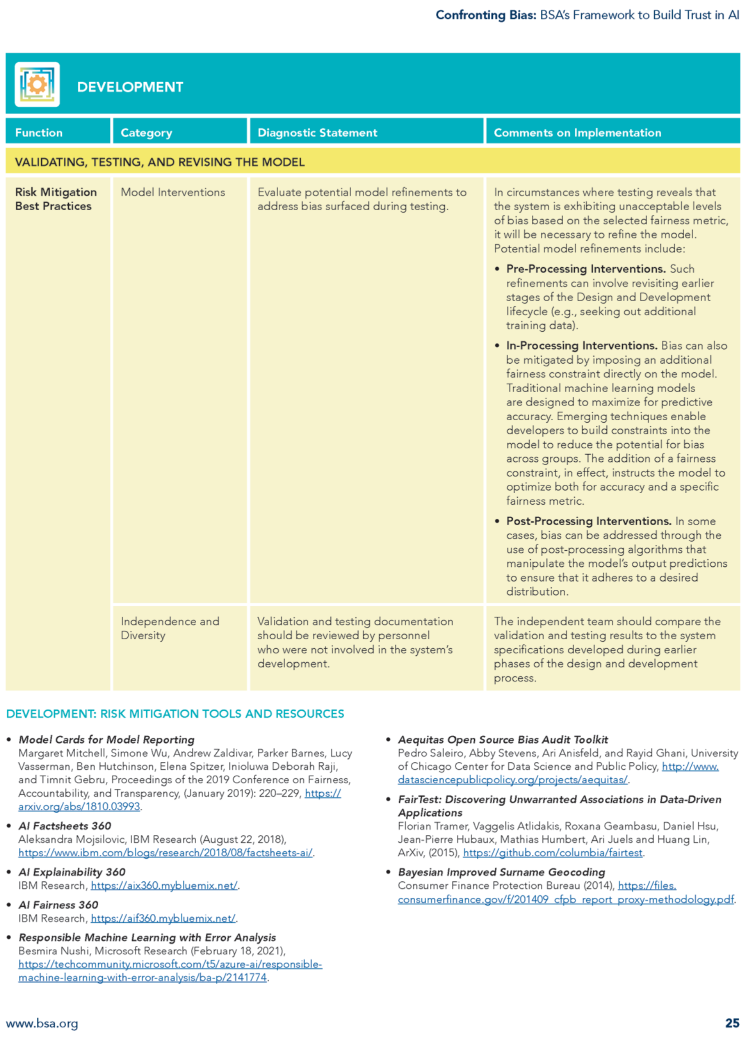
- ***Datasheets for Datasets***
  Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford, arXiv:1803.09010v7, (March 19, 2020), https://arxiv.org/abs/1803.09010.
- ***AI FactSheets 360***
  IBM Research, https://aif360.mybluemix.net/.

DEVELOPMENT

Function	Category	Diagnostic Statement	Comments on Implementation
DATA PREPARATION AND MODEL DEFINITION			
Impact Assessment	Document Feature Selection and Engineering Processes	Document rationale for choices made during the feature selection and engineering processes and evaluate their impact on model performance.	Examine whether feature selection or engineering choices may rely on implicitly biased assumptions.
		Document potential correlation between selected features and sensitive demographic attributes.	For features that closely correlate to a sensitive class, document the relevance to the target variable and the rationale for its inclusion in the model.
	Document Model Selection Process	Document rationale for the selected modeling approach.	
		Identify, document, and justify assumptions in the selected approach and potential resulting limitations.	
Risk Mitigation Best Practices	Feature Selection	Examine for biased proxy features.	• Simply avoiding the use of sensitive attributes as inputs to the system—an approach known as "fairness through unawareness"—is not an effective approach to mitigating the risk of bias. Even when sensitive characteristics are explicitly excluded from a model, other variables can act as proxies for those characteristics and introduce bias into the system. To avoid the risk of proxy bias, the AI Developer should examine the potential correlation between a model's features and protected traits and examine what role these proxy variables may be playing in the model's output. • The ability to examine statistical correlation between features and sensitive attributes may be constrained in circumstances where an AI Developer lacks access to sensitive attribute data and/or is prohibited from making inferences about such data.[22] In such circumstances, a more holistic analysis informed by domain experts may be necessary.

DEVELOPMENT

Function	Category	Diagnostic Statement	Comments on Implementation
DATA PREPARATION AND MODEL DEFINITION			
Risk Mitigation Best Practices *(continued)*	Feature Selection	Scrutinize features that correlate to sensitive attributes.	• Features that are known to correlate to a sensitive attribute should only be used if there is a strong logical relationship to the system's target variable. • For example, income—although correlated to gender—is reasonably related to a person's ability to pay back a loan. The use of income in an AI system designed to evaluate creditworthiness would therefore be justified. In contrast, the use of "shoe size"—which also correlates to gender—in a model for predicting creditworthiness would be an inappropriate use of a variable that closely correlates to a sensitive characteristic.
	Independence and Diversity	Seek feedback from diverse stakeholders with domain-specific expertise.	The feature engineering process should be informed by personnel with diverse lived experiences and expertise about the historical, legal, and social dimensions of the data being used to train the system.
	Model Selection	Avoid inscrutable models in circumstances where both the risk and potential impact of bias are high.	Using more interpretable models can mitigate the risks of unintended bias by making it easier to identify and mitigate problems.
VALIDATING, TESTING, AND REVISING THE MODEL			
Impact Assessment	Document Validation Processes	Document how the system (and individual components) will be validated to evaluate whether it is performing consistent with the design objectives and intended deployment scenarios.	
		Document re-validation processes.	• Establish cadence at which model will be regularly re-validated. • Establish performance benchmarks that will trigger out-of-cycle re-validation.
	Document Testing Processes	Test the system for bias by evaluating and documenting model performance.	Testing should incorporate fairness metrics identified during Design phase and examine the model's accuracy and error rates across demographic groups.
		Document how testing was performed, which fairness metrics were evaluated, and why those measures were selected.	
		Document model interventions.	If testing reveals unacceptable levels of bias, document efforts to refine the model.



DEVELOPMENT

Function	Category	Diagnostic Statement	Comments on Implementation
VALIDATING, TESTING, AND REVISING THE MODEL			
Risk Mitigation Best Practices	Model Interventions	Evaluate potential model refinements to address bias surfaced during testing.	In circumstances where testing reveals that the system is exhibiting unacceptable levels of bias based on the selected fairness metric, it will be necessary to refine the model. Potential model refinements include: • **Pre-Processing Interventions.** Such refinements can involve revisiting earlier stages of the Design and Development lifecycle (e.g., seeking out additional training data). • **In-Processing Interventions.** Bias can also be mitigated by imposing an additional fairness constraint directly on the model. Traditional machine learning models are designed to maximize for predictive accuracy. Emerging techniques enable developers to build constraints into the model to reduce the potential for bias across groups. The addition of a fairness constraint, in effect, instructs the model to optimize both for accuracy and a specific fairness metric. • **Post-Processing Interventions.** In some cases, bias can be addressed through the use of post-processing algorithms that manipulate the model's output predictions to ensure that it adheres to a desired distribution.
	Independence and Diversity	Validation and testing documentation should be reviewed by personnel who were not involved in the system's development.	The independent team should compare the validation and testing results to the system specifications developed during earlier phases of the design and development process.

DEVELOPMENT: RISK MITIGATION TOOLS AND RESOURCES

- ***Model Cards for Model Reporting***
  Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru, Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, (January 2019): 220–229, https://arxiv.org/abs/1810.03993.
- ***AI Factsheets 360***
  Aleksandra Mojsilovic, IBM Research (August 22, 2018), https://www.ibm.com/blogs/research/2018/08/factsheets-ai/.
- ***AI Explainability 360***
  IBM Research, https://aix360.mybluemix.net/.
- ***AI Fairness 360***
  IBM Research, https://aif360.mybluemix.net/.
- ***Responsible Machine Learning with Error Analysis***
  Besmira Nushi, Microsoft Research (February 18, 2021), https://techcommunity.microsoft.com/t5/azure-ai/responsible-machine-learning-with-error-analysis/ba-p/2141774.
- ***Aequitas Open Source Bias Audit Toolkit***
  Pedro Saleiro, Abby Stevens, Ari Anisfeld, and Rayid Ghani, University of Chicago Center for Data Science and Public Policy, http://www.datasciencepublicpolicy.org/projects/aequitas/.
- ***FairTest: Discovering Unwarranted Associations in Data-Driven Applications***
  Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels and Huang Lin, ArXiv, (2015), https://github.com/columbia/fairtest.
- ***Bayesian Improved Surname Geocoding***
  Consumer Finance Protection Bureau (2014), https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf.

DEPLOYMENT AND USE

Function	Category	Diagnostic Statement	Comments on Implementation
PREPARING FOR DEPLOYMENT AND USE			
Impact Assessment	Document Lines of Responsibility	Define and document who is responsible for the system's outputs and the outcomes they may lead to, including details about how a system's decisions can be reviewed if necessary.	
		Establish management plans for responding to potential incidents or reports of system errors.	• What does it mean for the system to fail and who might be harmed by a failure? • How will failures be detected? • Who will respond to failures when they are detected? • Can the system be safely disabled? • Are there appropriate plans for continuity of critical functions?
	Document Processes for Monitoring Data	Document what processes and metrics will be used to evaluate whether production data (i.e., input data the system encounters during deployment) differs materially from training data.	
	Document Processes for Monitoring Model Performance	For static models, document how performance levels and classes of error will be monitored over time and benchmarks that will trigger review.	
		For models that are intended to evolve over time, document how changes will be inventoried; if, when, and how versions will be captured and managed; and how performance levels will be monitored (e.g., cadence of scheduled reviews, performance indicators that may trigger out-of-cycle review).	
	Document Audit and End-of-Life Processes	Document the cadence at which impact assessment evaluations will be audited to evaluate whether risk mitigation controls remain fit for purpose.	
		Document expected timeline that system support will be provided and processes for decommissioning system in event that it falls below reasonable performance thresholds.	
Risk Mitigation Best Practices	Monitoring for Drift and Model Degradation	Input data encountered during deployment can be evaluated against a statistical representation of the system's training data to evaluate the potential for data drift (i.e., material differences between the training data and deployment data that can degrade model performance).	

DEPLOYMENT AND USE

Function	Category	Diagnostic Statement	Comments on Implementation
PREPARING FOR DEPLOYMENT AND USE			
Risk Mitigation Best Practices *(continued)*	Product Features and User Interface	Integrate product and user interface features to mitigate risk of foreseeable unintended uses—e.g., interface that enforces human-in-the-loop requirements, alerts to notify when a system is being misused.	
	System Documentation	AI Developers should provide sufficient documentation regarding system capabilities, specifications, limitations, and intended uses to enable AI Deployers to perform independent impact assessment concerning deployment risks.	If necessary, AI Developers can also provide AI Deployers with a technical environment to perform an independent impact assessment.
		Consider incorporating terms into the End-User License Agreement that set forth limitations designed to prevent foreseeable misuses (e.g., contractual obligations to ensure end-user will comply with acceptable use policy).	
		Sales and marketing materials should be closely reviewed to ensure that they are consistent with the system's actual capabilities.	
	AI User Training	AI Deployers should provide training for AI Users regarding a system's capabilities and limitations, and how outputs should be evaluated and integrated into a workflow.	For human-in-the-loop oversight of AI system to be an effective risk mitigation measure, AI Users should be provided adequate information and training so they can understand how the system is operating and make sense of the model's outputs.
	Incident Response and Feedback Mechanisms	AI Deployers should maintain a feedback mechanism to enable AI Users and Affected Individuals (i.e., members of the public that may interact with the system) to report concerns about the operation of a system.	For consequential decisions, Affected Individuals should be provided with an appeal mechanism.

DEPLOYMENT AND USE: RISK MITIGATION TOOLS AND RESOURCES

- ***AI Incident Response Checklist***
  BNH.AI, https://www.bnh.ai/public-resources.
- ***Watson OpenScale***
  IBM, https://www.ibm.com/cloud/watson-openscale.
- ***Detect Data Drift on Datasets***
  Microsoft Azure Machine Learning (June 25, 2020), https://docs.microsoft.com/en-us/azure/machine-learning/how-to-monitor-datasets?tabs=python#create-dataset-monitors.

Foundational Resources

A Framework for Understanding Unintended Consequences of Machine Learning
Harini Suresh and John V. Guttag, arXiv (February 2020), https://arxiv.org/abs/1901.10002.

AI Fairness
Trisha Mahoney, Kush R. Varshney, and Michael Hind, O'Reilly (April 2020), https://www.oreilly.com/library/view/ai-fairness/9781492077664/.

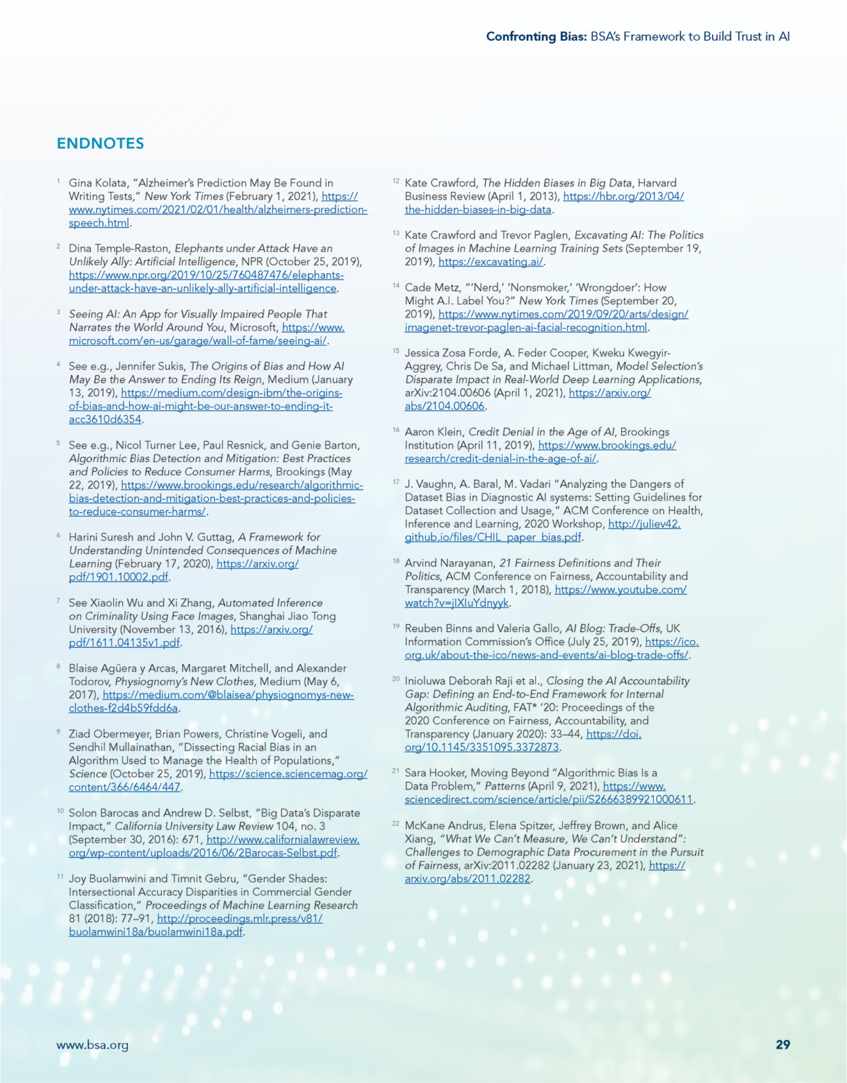
Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models
Andrew Burt, Brenda Leong, Stuart Shirrell, and Xiangnong (George) Wang, Future of Privacy Forum (June 2018), https://fpf.org/wp-content/uploads/2018/06/Beyond-Explainability.pdf.

Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI
Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach, CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (April 2020): 1–14, https://doi.org/10.1145/3313831.3376445.

Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing
Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P., FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, (January 2020): 33–44, https://doi.org/10.1145/3351095.3372873.

Supervisory Guidance on Model Risk Management
US Federal Reserve Board (April 2011), https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf.

Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector
David Leslie, The Alan Turing Institute (2019), https://doi.org/10.5281/zenodo.3240529.



ENDNOTES

[1] Gina Kolata, "Alzheimer's Prediction May Be Found in Writing Tests," *New York Times* (February 1, 2021), https://www.nytimes.com/2021/02/01/health/alzheimers-prediction-speech.html.

[2] Dina Temple-Raston, *Elephants under Attack Have an Unlikely Ally: Artificial Intelligence*, NPR (October 25, 2019), https://www.npr.org/2019/10/25/760487476/elephants-under-attack-have-an-unlikely-ally-artificial-intelligence.

[3] *Seeing AI: An App for Visually Impaired People That Narrates the World Around You*, Microsoft, https://www.microsoft.com/en-us/garage/wall-of-fame/seeing-ai/.

[4] See e.g., Jennifer Sukis, *The Origins of Bias and How AI May Be the Answer to Ending Its Reign*, Medium (January 13, 2019), https://medium.com/design-ibm/the-origins-of-bias-and-how-ai-might-be-our-answer-to-ending-it-acc3610d6354.

[5] See e.g., Nicol Turner Lee, Paul Resnick, and Genie Barton, *Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms*, Brookings (May 22, 2019), https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.

[6] Harini Suresh and John V. Guttag, *A Framework for Understanding Unintended Consequences of Machine Learning* (February 17, 2020), https://arxiv.org/pdf/1901.10002.pdf.

[7] See Xiaolin Wu and Xi Zhang, *Automated Inference on Criminality Using Face Images*, Shanghai Jiao Tong University (November 13, 2016), https://arxiv.org/pdf/1611.04135v1.pdf.

[8] Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov, *Physiognomy's New Clothes*, Medium (May 6, 2017), https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a.

[9] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science* (October 25, 2019), https://science.sciencemag.org/content/366/6464/447.

[10] Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," *California University Law Review* 104, no. 3 (September 30, 2016): 671, http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf.

[11] Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* 81 (2018): 77–91, http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

[12] Kate Crawford, *The Hidden Biases in Big Data*, Harvard Business Review (April 1, 2013), https://hbr.org/2013/04/the-hidden-biases-in-big-data.

[13] Kate Crawford and Trevor Paglen, *Excavating AI: The Politics of Images in Machine Learning Training Sets* (September 19, 2019), https://excavating.ai/.

[14] Cade Metz, "'Nerd,' 'Nonsmoker,' 'Wrongdoer': How Might A.I. Label You?" *New York Times* (September 20, 2019), https://www.nytimes.com/2019/09/20/arts/design/imagenet-trevor-paglen-ai-facial-recognition.html.

[15] Jessica Zosa Forde, A. Feder Cooper, Kweku Kwegyir-Aggrey, Chris De Sa, and Michael Littman, *Model Selection's Disparate Impact in Real-World Deep Learning Applications*, arXiv:2104.00606 (April 1, 2021), https://arxiv.org/abs/2104.00606.

[16] Aaron Klein, *Credit Denial in the Age of AI*, Brookings Institution (April 11, 2019), https://www.brookings.edu/research/credit-denial-in-the-age-of-ai/.

[17] J. Vaughn, A. Baral, M. Vadari "Analyzing the Dangers of Dataset Bias in Diagnostic AI systems: Setting Guidelines for Dataset Collection and Usage," ACM Conference on Health, Inference and Learning, 2020 Workshop, http://juliev42.github.io/files/CHIL_paper_bias.pdf.

[18] Arvind Narayanan, *21 Fairness Definitions and Their Politics*, ACM Conference on Fairness, Accountability and Transparency (March 1, 2018), https://www.youtube.com/watch?v=jIXIuYdnyyk.

[19] Reuben Binns and Valeria Gallo, *AI Blog: Trade-Offs*, UK Information Commission's Office (July 25, 2019), https://ico.org.uk/about-the-ico/news-and-events/ai-blog-trade-offs/.

[20] Inioluwa Deborah Raji et al., *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*, FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (January 2020): 33–44, https://doi.org/10.1145/3351095.3372873.

[21] Sara Hooker, Moving Beyond "Algorithmic Bias Is a Data Problem," *Patterns* (April 9, 2021), https://www.sciencedirect.com/science/article/pii/S2666389921000611.

[22] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang, *"What We Can't Measure, We Can't Understand": Challenges to Demographic Data Procurement in the Pursuit of Fairness*, arXiv:2011.02282 (January 23, 2021), https://arxiv.org/abs/2011.02282.

The Software Alliance

BSA

www.bsa.org

BSA Worldwide Headquarters
20 F Street, NW
Suite 800
Washington, DC 20001

☎ +1.202.872.5500
@BSAnews
@BSATheSoftwareAlliance

BSA Asia-Pacific
300 Beach Road
#30-06 The Concourse
Singapore 199555

☎ +65.6292.2072

BSA Europe, Middle East & Africa
44 Avenue des Arts
Brussels 1040
Belgium

☎ +32.2.274.13.10

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

Good afternoon Chairman Foster, Ranking Member Gonzalez, and Members of the AI Task Force. My name is Meg King. I am Director of the Science and Technology Innovation Program at the Woodrow Wilson International Center for Scholars, a nonpartisan think tank created by Congress nearly sixty years ago.

My program studies the policy opportunities and challenges of emerging technologies, investigates opportunities to foster more open science and builds serious games. We also offer hands-on training programs – called the Technology Labs – to Congressional and Executive branch staff on a variety of issues including artificial intelligence. Next month, we will offer a series of individual trainings on AI for Members as well. Thank you for inviting me to testify today.

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 have seen AI function spectacularly – 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.

EQUAI

Testimony of
Miriam Vogel
President & CEO, EqualAI

Hearing entitled: "Beyond I, Robot: Ethics, Artificial Intelligence, and the Digital Age"

Before the Task Force on Artificial Intelligence
United States House Committee on Financial Services
October 13, 2021

Chairman Foster, Ranking Member Gonzalez, and distinguished members of the Task Force on Artificial Intelligence (AI), thank you for conducting this critical hearing and for the opportunity to submit this testimony. The work you have done in past hearings has been important to clarify and understand the issues and challenges surrounding AI development and use, and I commend you for delving into this next question of the ethical implications of AI.

My name is Miriam Vogel, and I am the President and CEO of EqualAI, a nonprofit organization that was founded with the express purpose of reducing unconscious bias in AI systems. At EqualAI, we are AI net-positive. We believe that AI is and will be a powerful tool to advance our lives, our economy, and our opportunities to thrive. But only if we are careful to ensure that the AI we use does not perpetuate and mass produce historical, and new unanticipated forms, of biases and discrimination.

I was asked to lead this organization with a background that is currently unconventional in this space but hopefully will become more common as we invite collaborative, multi-stakeholder efforts. I worked at the intersection of technology, policy and the law as a lawyer in private practice and in government. I previously had the opportunity to address the problem of bias in more traditional contexts, such as leading President Obama's Equal Pay Task Force and the effort to create implicit bias training for federal law enforcement at the Department of Justice under the direction of Deputy Attorney General Sally Yates. Given that orientation, our focus at EqualAI is on driving multi-staker efforts to ensure our technology platforms are equitable and inclusive. We perceive implicit bias in AI as an age-old issue that is surfacing in a new medium, but now at scale and with graver potential impacts.

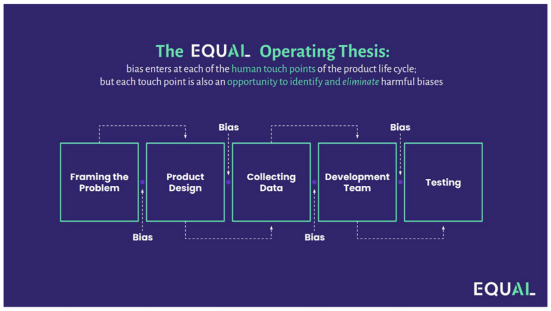
We believe we are at a critical juncture because AI is becoming an increasingly important part of our daily lives - while decades of progress made and lives lost to promote equal opportunity can essentially be unwritten in a few lines of code. And the perpetrators of this disparity may not even realize the harm they are causing. Our country's long history of housing discrimination is being replicated at scale in mortgage approval algorithms that determine credit worthiness using proxies for race and class. An exciting innovation in AI deep learning language modeling, GPT-3, is also demonstrating its problematic biases, such as generating stories depicting sexual encounters involving children and exhibiting biases against people based on their religion, race and gender. Our goal at EqualAI is to help avoid perpetuating these harms by offering programs, frameworks and strategies to establish responsible AI governance.

We focus our efforts at EqualAI on supporting three main stakeholders: companies, policymakers and lawyers. We work with companies to help them address and reduce the infusion of implicit bias in their AI systems. We aim to support policy makers in the essential task of establishing the appropriate guardrails that support innovation while mitigating harmful bias in AI. For instance, we look forward to supporting the important work underway at the National Institute of Standards and Technology (NIST) next week by moderating an AI Risk Management Framework (RMF) workshop panel. Finally, we bring lawyers into this effort. Lawyers need to help companies understand and manage the risks in the AI systems they are building, acquiring, using and deploying by ensuring they employ frameworks to reduce both harms and liabilities.

Often, the first step in our work is helping companies come to terms with the reality that they are now effectively AI companies. Two decades ago, most companies did not realize that they needed to have cohesive plans and contingencies in place to protect against the unauthorized exploitation of systems, networks and technologies. Today, however, companies widely recognize the importance of cybersecurity. Much like the trajectory of cybersecurity awareness, companies now need to adjust and understand that they use AI in one or more pivotal functions-hiring, credit lending, health care determinations, to name a few- and must, therefore, have a governance plan in place to address the potential discrimination that these systems can dispense at scale. This is of particular concern with AI given that key determinations and assessments occur behind the thick veil of the proverbial 'black box', where the algorithm's inputs and operations are generally unknown to the end user. We support businesses' efforts to establish AI governance and best practices in a variety of ways. For instance, we just launched a badge program, in collaboration with the World Economic Forum, to train senior executives on how to understand and implement responsible AI governance and create a community comprised of companies and individuals committed to this effort.

Implicit Bias in AI

Your past hearings provided an important overview of both the benefits and risks of our accelerated use of AI in the financial and housing sectors. One grave concern that has been repeatedly articulated is the infusion and dissemination of implicit bias at scale through AI systems. Implicit or unconscious bias is based on a stereotype, or characterization of people of a certain group which can be positive or negative. It is noticing patterns and making generalizations based on those assumptions. As referenced in your past hearings, implicit bias embeds in AI in a variety of ways. Our operating thesis is that bias can embed in each of the human touch points throughout the lifecycle of the creation of an AI system (see *Diagram 1* below). From the ideation phase- deciding which problem you want to use AI to help solve, to the design, data collection, development and testing phases. Each stage is limited by the experience and imagination of those on that team, which is reinforced by historical and learned biases in the data. But we are also optimistic and think each touchpoint is an opportunity to identify and eliminate harmful biases. As such, risk management should occur at each stage of the AI lifecycle.



The EQUAI Operating Thesis:

bias enters at each of the human touch points of the product life cycle;
but each touch point is also an opportunity to identify and *eliminate* harmful biases

Framing the Problem

Product Design

Collecting Data

Development Team

Testing

Bias

Bias

Bias

Bias

EQUAI

Diagram 1

Framework to Combat Implicit, Harmful Bias in AI

There are an increasing number of frameworks that provide helpful guidance on methods to identify and reduce harms from AI systems before they materialize (e.g. GAO Framework; GSA Center of Excellence Guide to AI Ethics, DoD Ethical Principles for AI, AI Ethics Framework for the Intelligence Community). There are also efforts underway to clarify and standardize frameworks and best practices, such as the important work at NIST to support AI standards development, develop a risk management framework for trustworthy AI systems, and develop best practices for documenting and sharing data sets used to train AI systems, pursuant to the National Defense Authorization Act for Fiscal Year 2021 (NDAA).

We offer an EqualAI Framework (see *Diagram 2* below) as a general guide with five “pillars” a company should consider as part of its effort to establish enterprise-wide responsible AI governance. These recommendations are particularly important in sensitive sectors such as finance and housing, given the real potential for biased AI to perpetuate discrimination against job candidates, renters, mortgage seekers, insurance applicants, and disadvantaged small businesses seeking capital. These “pillars” are as follows:

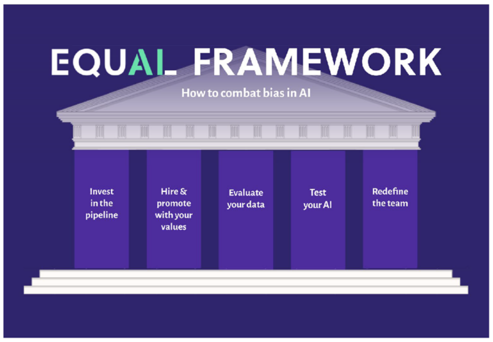


Diagram 2

1. *Invest in the pipeline*

Artificial intelligence needs to be created by and for a broader cross-section of our population. Research suggests that homogeneous teams – like those that comprise many of the teams coding our AI – are more likely to generate biased algorithms than diverse teams. We have also seen that lack of diversity in AI creation could rise to a life or death issue given the ultimate uses for many of these technologies, such as determining who can access ventilators and critical health care services during a crisis or deciding an individual's fate in the criminal justice system.

Many organizations are ensuring that our next generation of coders and tech executives represent a broader cross-section of our population (e.g., AI4All, Black Girls Code, Girls Who Code, Code.org, etc.). Their work is critical to ensuring that our AI benefits from broader perspectives and that more communities thrive in the AI economy.

We currently offer inadequate access to computer science and engineering courses in our classrooms. Reports indicate that only 22 states have K-12 standards for computer science education, and only 15 require high school computer science courses. Where these courses are offered, they are often rife with race, class, geographic and gender imbalances, depriving our workforce, and resulting AI, of the full breadth of American talent.

Congress can play a key role by offering funding for teachers to learn and teach coding as part of the K-12 curriculum in our nation's public schools. Congress can also direct additional funding to ethical AI research and development in our higher education system, including at HBCUs, minority-serving institutions, and community colleges.

2. Hire and promote with your values

To create and sustain a diverse workplace, and produce better AI, employees. and managers in particular, must be trained to recognize and address implicit bias in human resource functions (e.g., hiring, evaluation, promotion and termination decisions). AI programs used in these human resource functions should be checked to ensure they are in sync with a company's values. AI used for hiring, evaluations, promotions and terminations could be infused with bias and as such, must be checked- and constantly rechecked- for harmful biases given the likelihood it will constantly learn new patterns and may offer inequitable employment decisions. A few best practices we recommend at this stage include: (1) ensuring a broad cross-section of diversity in each candidate panel, (2) keeping humans involved and 'in the loop' in decision-making processes and (3) constantly checking for biased patterns or outcomes with simulated or hypothetical personas.

Congress can help by clarifying that hiring and employment laws prohibiting discrimination are equally applicable to AI-driven or supported recommendations.

---

3. Evaluate your data

The more we know about the datasets on which AI is built and trained, the safer we are as a society. We encourage companies to identify the gaps in the data so that they can be rectified or at a minimum, clarified for end users. We offer the EqualAI Checklist© as a starting place to evaluate data sets and identify possible liabilities. There are other helpful resources that provide best practices to ensure data sources, as well as gaps, over- and under-representations, are identified, such as Federal Trade Commission (FTC) guidance. Best practices include seeking to answer: (1) How representative is the data set? (2) Does the data model account for biases? (3) How accurate are the predictions that the AI offers? (4) For whom could this system fail?

Congress can play a key role by clarifying expectations and specifying the information that should be provided when brokering data sets or AI training data, similar to current expectations for nutrition labels. Industry, and society as a whole, would benefit from notice and clarity on what data points should be highlighted and particularly by those selling or sharing the data sets to help with predictive modeling and other AI uses. This can help ensure uniform standards and identify gaps that need to be addressed when using the data that would otherwise be unknown or unspecified, leading to potentially harmful and inequitable outcomes.

In the financial services context, intensive data evaluation can also help ensure the use of AI does not result in violations of existing statutes. Laws such as the Fair Housing Act and the Equal Credit Opportunity Act are critical, albeit imperfect, safeguards against discrimination for those seeking to build equity, access capital, and pursue better opportunities for themselves and their families. Congressional vigilance is critical to ensuring that these safeguards are not eroded given that creditworthiness, underwriting, and other decision-making processes are increasingly automated. These hearings by the Task Force on AI are important steps in attaining this goal.

---

4. Test your AI

AI systems, and particularly those that are customer-facing or used in human resource functions, should be checked for bias on a routine basis. Given that AI constantly iterates and learns new patterns as it is fed new data, it will often adopt new biases. Good AI governance includes routine audits and checks to ensure recommendations are consistent with expectations and that outlier outcomes are investigated.

Responsible AI governance includes:

- identifying which values and biases will be tested routinely;
- articulating the stages of the AI lifecycle development at which the testing will conducted (e.g., pre-design, design and development, deployment);
- establishing the cadence for testing;
- documenting relevant findings and the completion of each stage to promote consistency, accountability and transparency; and
- identifying the designated point of contact who owns this responsibility ultimately, including: coordinating incoming questions and concerns, ensuring that responses are consistent and that new challenges are elevated and addressed appropriately.

The FTC guidance and EqualAI Checklist© are two sources for additional guidance and there is a growing body of experts and algorithmic auditors to help test AI systems. Congress can help normalize the practice of algorithmic audits and by collecting information about a company's high level AI governance plan and the appropriate point person. The submission of this information could be made a common practice as part of routine filings, such as with the Departments of Housing and Labor, Office of the Comptroller of the Currency (OCC), Consumer Financial Protection Bureau (CFPB), FTC or Securities and Exchange Commission (SEC).

---

5. Redefine the team

AI products should be tested prior to their public release by and for those under-represented in its creation or in the data used to build and train the system. Special consideration should be given, and broader audiences should be brought in to help determine potential end users and those who could be impacted downstream who were insufficiently represented on the creation team and in the datasets used to build and train the program.

Congress can help support this important aspect of responsible AI governance by ensuring that it is following these and other best practices with its own internal AI development and procurement processes and by sharing best practices, resources, and lessons learned with industry and the general public. This Committee can help ensure that our nation's financial regulators, and the institutions they oversee, are leaders in this regard.

Additional Proposed Solutions

In addition to frameworks, there are numerous additional ways that Congress can play an instrumental role in ensuring more effective, less harmful AI development and deployment.

The National Security Commission on Artificial Intelligence (NSCAI), like prior cybersecurity reports, has warned that "America is not prepared to defend or compete in the AI era." The report recommends the establishment of the foundations for widespread integration of AI by 2025, including digital infrastructure and developing a digitally-literate workforce. These recommendations are critical to ensuring we have the critical mass of Americans necessary to perform AI-created and supported jobs. They also will enable us to support vulnerable and underrepresented populations that otherwise could fall subject to an even wider and more dangerous income disparity gap.

1. Auditing

We support mandates for auditable AI for systems used in pivotal functions, where AI systems are queried externally with hypothetical cases that are either synthetic or real. The more transparency the better, and in particular, notice of populations who are under or over represented in underlying datasets and for whom the AI system will have different success rates should be encouraged in the form of nutritional labels, as mentioned above. However, when this is not possible, and even when it is, there should be an expectation of routine, external audits with publicly available and easily accessible results.

The necessity of algorithmic auditing is particularly evident in the financial services context given that these AI-supported recommendations directly impact people's lives and opportunities and yet, are rooted in a part of our history, and thus trained on data, that is rife with imbalance and discrimination.

2. Incentivize investment in the Future of Work

There is a palpable concern that AI will edge humans out of the workplace, as addressed at a congressional hearing on the Future of Work in 2019. To be sure, automation, in tandem with the COVID-19 recession, is creating a 'double-disruption' scenario for workers. Like all transformative technologies, AI has and will inevitably eliminate jobs, but it will also open up possibilities, many of which we do not yet realize. Some estimate that by 2025, 85 million jobs may be lost but 97 million new roles may emerge due to automation in the workforce.

The U.S. is estimated to spend approximately 0.1% of its GDP on retraining programs, which is one-fifth of the average expenditure for countries in the Organization for Economic Cooperation and Development (OECD).

To lead in the AI revolution, safeguard our economy and support greater prosperity in more communities, we need to reskill our workforce. Some estimate that businesses could collectively reskill 45% of workers at risk of losing their jobs but, if governments join this effort, we could reskill as many as 77% of at risk workers. This would benefit government and society directly with increased tax returns and lower social costs, including reduced homelessness and food insecurity.

In addition, Congress could commission a study to better understand and articulate the type of skills and jobs that will likely emerge, enabling us to educate and upskill accordingly. One often cited study found that there are as many predictions of what the new jobs will look like as there are experts. The best way for us to plan for the future workforce is to offer clarity and evidence-based assessments of what it will look like. This study could be part of current, related efforts, such as the National Academies AI Impact Study on Workforce, (per Section 5105 of the NDAA), and the National AI Advisory Committee (per Section 5104(d) (4) of the NDAA).

An additional significant contribution could be the inclusion of additional tax breaks for investments in upskilling employees, loan forgiveness for graduates with computer science degrees who spend a minimum number of hours teaching in K-12 classes and increasing opportunities for secondment both for those with technical skills to support schools and government regulators and for government employees to spend time in the private sector.

3. *Bill of rights*

We need to ensure that the general public is empowered to require that AI systems, and other technologies we use, respect our democratic values and right to be free from discrimination. One such solution is the new “AI bill of rights” proposed last week by Dr. Eric Lander and Dr. Alondra Nelson of the White House Office of Science and Technology. It would ensure that the public is put on notice of critical information, such as when and how AI is influencing a decision that affects our civil rights and civil liberties and when we are using AI that has not been audited for implicit biases or trained on sufficiently representative data sets. Likewise, it envisions an opportunity for meaningful recourse for individuals harmed by such algorithms.

In conclusion, as we noted at the outset, we believe that it is imperative at this critical juncture to ensure that AI is built by and for a broader cross-section of our population. It is not only the right thing to do, a strong U.S. economy and our leadership depend on it.

Thank you again for the opportunity to testify before the Task Force. I look forward to answering your questions.

Testimony before US House of Representatives, Financial Services Committee, the Task Force on Artificial Intelligence

on *Beyond I, Robot: Ethics, Artificial Intelligence, and the Digital Age*

by Jeffery Yong, Principal Advisor, Financial Stability Institute, Bank for International Settlements

13 October 2021

Good afternoon Chair Foster, Ranking Member Gonzalez and distinguished members of the Task Force. My name is Jeffery Yong and I am a Principal Advisor at the Financial Stability Institute of the Bank for International Settlements (BIS). I offer my remarks today entirely in my personal capacity based on a publication that I co-authored with my colleague Jermy Prenio entitled FSI Insights no 35, *Humans keeping AI in check – emerging regulatory expectations in the financial sector.*[1] The views expressed in that paper are our own and do not necessarily represent those of the BIS, its members or the Basel-based committees. I am appearing before the Task Force voluntarily and would like to note that my statements here today are similarly my personal views, and they do not represent the official views of the BIS, its members or the Basel-based committees.

By way of background, the Financial Stability Institute (FSI)[2] is a unit within the BIS with a mandate to support implementation of global regulatory standards and sound supervisory practices by central banks and financial sector regulatory and supervisory authorities worldwide. One of the ways the FSI carries out this mandate is through its policy implementation work, which involves publishing FSI Insights papers. The papers aim to contribute to international discussions on a range of contemporary regulatory and supervisory policy issues and implementation challenges faced by financial sector authorities.

In preparing FSI Insights no 35, my co-author and I found that regulatory expectations on the use of artificial intelligence (AI) in financial services were at a nascent stage. Accordingly, we drafted the paper with four key objectives:

1. to identify emerging common financial regulatory themes surrounding AI governance;
2. to assess how similar or different these common regulatory themes are viewed in the context of AI vis-à-vis that of traditional financial models;
3. to explore how existing international financial regulatory standards may be applied in the context of AI governance; and
4. to examine challenges in implementing the common regulatory themes.

To this end, we canvassed a selection of policy documents on AI governance issued by financial authorities or groups formed by them, as well as other cross-industry AI governance guidance that apply to the financial sector. In total, we examined 19 policy documents issued by 16 national or regional authorities and two international organisations. Most of these documents are either discussion papers or high-level principles, which underscores the fact that financial regulatory thinking in this area is at a very early stage.

We identified five common themes that recur in the policy documents that we examined. These are reliability, accountability, transparency, fairness and ethics.

[1] See FSI Insights, no 35, *Humans keeping AI in check – emerging regulatory expectations in the financial sector*.

[2] See Financial Stability Institute.

On the theme of reliability, emerging supervisory expectations for AI and traditional models appear to be similar. What seems to be different is that the reliability of AI models is viewed from the perspective of avoiding harm to data subjects, for example through discrimination.

On the theme of accountability, it is acknowledged that both traditional and AI models require human intervention. In the case of AI, however, this requirement is motivated by the need to make sure that decisions based on AI models do not result in unfair or unethical outcomes. Moreover, external accountability is emphasised in the case of AI models so that data subjects are aware of AI-driven decisions and have channels for recourse.

On the theme of transparency, supervisory expectations related to explainability and auditability are similar for AI and traditional models. However, expectations on external disclosure are unique to AI models. This refers to expectations that firms using AI models should make data subjects aware of AI-driven decisions that impact them, including how their data is used.

On the theme of fairness, there is a distinct and strong emphasis in emerging supervisory expectations on this aspect in the case of AI models. Fairness is commonly described in the policy documents we reviewed in terms of avoiding discriminatory outcomes.

Similarly on the theme of ethics, there is also a distinct and strong emphasis on this aspect in AI models. Ethics expectations are broader than fairness and relate to ascertaining that customers will not be exploited or harmed, either through discrimination or other causes (eg AI using illegally obtained information).

Given the similarities of these themes in the context of AI and traditional financial models, existing financial regulatory standards that govern the use of traditional models may be applied in the context of AI. However, there may be scope to do more in defining financial regulatory expectations related to fairness and ethics. These could supplement consumer protection laws that cover non-discrimination clauses, which could also apply in the context of the use of AI in the financial sector.

The use of AI in the financial sector, however, presents certain challenges in a direct application of existing financial regulatory requirements. A key challenge is due to the level of complexity and lack of explainability that characterise AI models. These limit the transparency of the models, which in turn makes it challenging for financial supervisors to assess the reliability, accountability, fairness and ethics in the use of AI in the financial services industry.

A way to overcome these challenges is to consider a tailored and coordinated regulatory and supervisory approach. This means differentiating the regulatory and supervisory treatment on the use of AI models, depending on the conduct and prudential risks that they pose. In addition, coordination between conduct, prudential, as well as data protection authorities will help address the cross-cutting implications of the use of AI models in the financial services industry.

Thank you very much.

House Financial Services Task Force on Artificial Intelligence hearing: “Task Force on Artificial Intelligence: Beyond I, Robot: Ethics, Artificial Intelligence, and the Digital Age” on October 13, 2021

Please find below my reply to the questions from Congresswoman Sylvia R. Garcia. I am responding voluntarily in my personal capacity and my views are my own and do not represent those of the BIS, its members or the Basel-based committees.

1. Challenges relating to transparency and fairness

A key challenge in meeting the transparency goals of AI models is due to the complexity and lack of explainability of certain types of machine learning (ML) algorithms. It is challenging for financial institutions to explain complex ML algorithms in a way that can be understood by regulators. Some ML algorithms such as deep learning and neural networks are considered as ‘black box’, which produce sensible results but are difficult to explain or proof unlike traditional statistical models. This is because such ML algorithms work through complex interactions between multiple variables, inferring attributes from data inputs and placing different weights on different data attributes. This also makes it difficult to disclose to data subjects, including financial consumers, in plain and simple language what data are used and how these affect the decision relevant to them.

Another challenge relates to determining the appropriate level of transparency of AI/ML models based on the target audience and materiality of the models’ results. In general, the more critical the use case is, the more important an algorithm should be transparent. Otherwise, the model’s results can be refuted and people accountable could lose trust in the model and refuse to take responsibility. Inadequate transparency could also erode consumer confidence or dissuade customers from using AI/ML-powered financial solutions. On the other hand, it should be pointed out that excessive transparency could create confusion or unintended opportunities for individuals including customers of financial institutions to exploit or manipulate AI/ML models.

It is important to note that transparency of an AI/ML algorithm is a pre-requisite to fulfilling some of the other sound AI governance principles. If a model is not transparent, it will be difficult to assess its reliability, performance and fairness apart from assessing the model's outcome against a specified benchmark. It will also be difficult to establish accountability if it is unclear which components of the algorithm are causing errors.

As regards fairness, a key challenge is the lack of universally accepted definitions of this term. Some regulators have left it to firms to come up with their own definitions but certain firms may find it difficult to do so. Even if they could, the definitions could fall short of general expectations by consumers.

In general, AI governance principles on fairness require human-in-the-loop or human-on-the-loop approaches. In practice, this means financial institutions need to allocate sufficient human resources in any AI implementation to ascertain fair and ethical results. In a way, there is some irony in having humans as safeguards to unfairness and unethical AI results when the latter is basically just reflecting these human flaws. Nevertheless, given that one of the main benefits of AI is to reduce the need for human intervention, it will be challenging to strike a right balance in fulfilling the human-in/on-the-loop expectations versus reaping the full benefits from technology automation.

2. Achieving end goals of transparency and fairness

There is an opportunity to build on the emerging common themes on AI governance in the financial sector as I explained in my opening remarks at the hearing. Financial authorities' views on how these common themes should be implemented are still evolving. A continued exchange of views and experiences at the international level may eventually lead to the development of guidance or standards, which could be helpful particularly to jurisdictions that are starting their digital transformation journey. Standards can also serve as a minimum benchmark to guide an orderly deployment of AI technologies within the financial sector. As more specific regulatory approaches or supervisory expectations emerge on specific aspects of AI use cases, there could be further work to identify such common "best practice" that will be useful for other jurisdictions to consider. At the same time, given the rapidly evolving technology trends, principles-based guidance continue to have its benefits and can complement such best practice approach.

The challenges and complexity presented by AI call for a tailored and coordinated regulatory and supervisory response based on the AI model's implications for conduct and prudential risks. The more AI model's use can potentially impact authorities' conduct and prudential objectives, the more stringent the relevant reliability/soundness, accountability, transparency, fairness and ethics requirements should be. For example, AI models used for credit underwriting decisions might be subjected to higher expectations than those used for customer support chatbots. In terms of transparency specifically, in practice this could mean that AI models for credit underwriting would be subject to more stringent disclosure requirements, eg akin to the legal requirement in the US that credit reporting agencies should disclose to consumers, upon request, all information that goes into their credit score, and to have them fixed if there are mistakes.

It should also be acknowledged that the use of AI by financial institutions will have implications for profitability, market impact, consumer protection and reputation. This calls for more coordination between prudential and conduct authorities in overseeing the deployment of AI in financial services.

While existing standards, laws and guidance may be used to address most AI-related issues, there may be scope to do more when it comes to fairness. Given that most of the issues associated with the use of AI are similar to those for traditional models, authorities can leverage existing standards, laws and guidance intended for the latter in assessing the former. Fairness, particularly as it relates to avoiding discriminatory outcomes, while identified as very important in the case of AI, may not be explicit in consumer protection laws in some jurisdictions. Making non-discrimination objectives explicit may help provide a good foundation for defining fairness in the context of AI, provide a legal basis for financial authorities to issue AI-related guidance and, at the same time, ensure that AI-driven, traditional model-driven and human-driven decisions in financial services are assessed against the same standard.
Jeffery Yong