Subcommittee Chair Stevens, Subcommittee Ranking Member Feenstra, Committee Chair Johnson, Ranking Member Lucas, and distinguished members of the subcommittee, my name is Dr. Charles Isbell and I am a Professor in and Dean for the College of Computing at Georgia Tech. Thank you for the opportunity to appear before this Subcommittee to discuss:

1. The importance of a culture of responsibility around artificial intelligence (AI) systems.
2. The need for transparency in AI systems in order to identify harmful bias.
3. Mitigation of the risks in AI.

By way of explaining my background, let me note that while I tend to focus on statistical machine learning, my research passion is actually artificial intelligence. I like to build large integrated systems, so I also tend to spend a great deal of my time doing research on autonomous agents, interactive entertainment, some aspects of human-computer interaction, software engineering, and even programming languages

I think of my field as interactive artificial intelligence. My fundamental research goal is to understand how to build autonomous agents that must live and interact with large numbers of other intelligent agents, some of whom may be human. Progress towards this goal means that we can build artificial systems that work with humans to accomplish tasks more effectively; can respond more robustly to changes in environment, relationships, and goals; and can better co-exist with humans as long-lived partners.

As the members of this Subcommittee well know, there has been an explosion in the development and deployment of what we might call AI technology. With that explosion has come a corresponding explosion in interest in AI.

In any discussion—particularly technical ones—it helps to define our terms. There are many potential definitions of AI. My favorite one is that it is “the art and science of making computers act like they do in the movies.” In the movies, computers are often semi-magical and anthropomorphic; they do things that, if humans did them, we would say they required intelligence.

This definition is borne out in our use of AI in the everyday world. We use the infrastructure of AI to search billions upon billions of documents to find the answers to a staggering variety of questions—often expressed literally as questions. We use automatically tagged images to organize our photos, and we use that same infrastructure to plan optimal routes for trips—even altering our routes on-the-fly in the face of changes in traffic. We are able to automatically detect tumors from x-rays, even those that trained doctors find difficult to see. We let computers finish our sentences as we type texts and use search engines, sometimes facilitating a subtle shift from prediction of our behavior to influence over our behavior. Often we take advantage of these services by using our phones (our phones!) to interpret a wide variety of spoken commands.

So, in some very important sense, AI already exists. It is not the AI of science fiction, neither benevolent intelligences working with humans as we traverse the galaxy, nor malevolent AI that seeks humanity's destruction. Nonetheless, we are living every day with machines that make decisions that, if humans made them, we would attribute to intelligence. And the machines often make those decisions faster and better than humans would.

Importantly, each of the examples we consider above is a distinctly human-centered problem. It is human-centered both in the sense that these systems are trying to solve problems that humans deal with every day—question answering, symptom evaluation, navigation—but also human-centered in the sense that humans have or currently perform some of those tasks. Presumably, these developments are all to the good. We are living up to the promise of technology that allows us to automate away work that is dirty, dangerous, or dull, freeing up human capital to be more productive, and, hopefully, for humans to be more fulfilled. The social and economic benefits are potentially immense.

There are also some reasons for concern. Those who work in the field will tell you that very often they aren't sure exactly how their algorithms reach the correct answer, only that they do. AI scientists describe these algorithms as "black box models."

The second concern is that sometimes those algorithms reach the wrong conclusion, and in a way that harms people and society. Artificial intelligence has all too often automated the biases of its programmers, or baked into its data. As a result, AI products have already been caught making biased decisions in banking, hiring, health care and criminal justice.

For example, according to the Marshall Project, almost every state uses some form of "risk assessment" at some stage in the criminal justice system.

> Risk assessments have existed in various forms for a century, but over the past two decades, they have spread through the American justice system, driven by advances in social science. The tools try to predict recidivism — repeat offending or breaking the rules of probation or parole — using statistical probabilities based on factors such as age, employment history, and prior criminal record. They are now used at some stage of the criminal justice process in nearly every state. Many court systems use the tools to guide decisions about which prisoners to release on parole, for example, and risk assessments are becoming increasingly popular as a way to help set bail for inmates awaiting trial.

This automated process relies on an algorithm in lieu of a judge's discretion. As noted by Cathy O'Neil, author of *Weapons of Math Destruction*, the data used by these algorithms to build models are sometimes suspect. Worse, we treat the output as "objective" without understanding that the data are themselves not objective. In this particular case, we set out to predict recidivism as if that means *the chance of committing a crime again* when in fact we are predicting *the chance of being arrested and convicted again.*

It does not take much imagination to see how being from a heavily policed area raises the chances of being arrested again, being convicted again, and in aggregate leads to even more policing of the same areas, creating a feedback loop. One can imagine similar issues with determining fit for a job, or credit-worthiness, or even face recognition and automated driving. In computing, we call this garbage-in-garbage-out: an algorithm is only as good as its data. This saying is certainly true, and especially relevant for AI algorithms that learn based on the data they are given.

Luckily, one way to address these issues is straightforward: to increase transparency. The kind of data the algorithm uses to build its model should be available. The decisions that such algorithms make should be inspectable. In other words, as we deploy these algorithms, each algorithm should be able to explain its output. "This applicant was assigned high risk because…" is more useful than, "This applicant was assigned high risk."

If algorithms are inspectable, their creators are then able to call in outside experts to inspect them. After all, those with the knowledge to design an artificial intelligence algorithm can't be expected to also be experts in medicine, the law, criminal justice, or banking. And outside experts shouldn't have to get a Ph.D. in computer science to understand what programmers are doing with their data and their theories. AI transparency allows for a much wider range of input into any given project. And when things go wrong, it shows exactly where and how.

The idea of AI transparency is straightforward, but its implementation will be more complicated. First, the complexity of the algorithms makes it impractical for humans to inspect them manually. We will need tools that translate the complexity of AI algorithms into useable human-scaled insights.

Second, researchers have demonstrated that the more transparent an AI is, the easier it is to hack. Or worse still, if the AI is a trade secret, the easier it is to replicate. Therefore, we will also need new tools to secure every part of the programming and training process from unwanted intruders.

This does not mean that transparent AI is impossible, just that it presents a series of important technical challenges. But we must also recognize that transparency isn't the only measure we can and should be taking to make AI responsible.

We also have the responsibility to consider the data sets that are used to train these algorithms. As shown in the earlier example about risk assessment for parolees, sometimes the data is skewed by the method that was used to collect it. This is a common problem in algorithms trained on social media data, to give another example.

Sometimes, the data set simply doesn't contain enough information about underrepresented groups to even recognize them as a group. If that is the case, the data set can be expanded to include more information about those groups. Alternatively, they can add another "learner"

program to the AI that focuses on identifying those groups. This in and of itself presents a considerable challenge, however, because it suggests that the only way to make systems more responsible is to make them more complicated. To solve that problem, we need new concepts in computing theory to help us organize responsible AIs more efficiently. There is precedent for putting practice before theory; people wrote in code for thousands of years before the theory underlying modern public-key cryptography was laid out in the 1970s.

These technical problems present some of the major research challenges in artificial intelligence today. The National Institute of Standards and Technology's ongoing effort to create an AI risk management framework will need to incorporate these technical questions and others.

There are, of course, human issues as well. Right now, about 66 percent of tech workers are white, and 20 percent are Asian. Roughly 75 percent are men. Now, I work in AI, and I am not alleging that my colleagues are racist or misogynist. I am pointing out, however, that people from a subset of the population often build products that affect everyone. And often, they don't realize they're missing valuable perspectives.

In the long term, one of the key solutions to AI bias will be bringing a wider group of people into computing education, and into machine learning more specifically. We need to improve both the number and the diversity of people entering the field, starting from K-12 and extending to post-graduate work. One major obstacle is a lack of instructors at every level. In my own state, Georgia, only 35 percent of high schools that have AP programs offer AP Computer Science.

Now, K-12 isn't the only place for intervention, and programming is not the only job in artificial intelligence. In my own college, our DataWorks program trains unemployed adults to clean and integrate data sets for use in artificial intelligence projects. There are opportunities to open AI careers to more communities at every point in the pipeline.

While technical solutions are important, as are diversity and equity, a larger culture change is also needed. Computing has long been an intellectual Wild West, where things changed so fast that the priority was always to find the next, better solution. Now, we have succeeded in finding solutions so good that they are entwined in nearly every area of our personal lives and communities.

We have not as a field caught up to the reality of that responsibility. Unlike engineers or lawyers or medical professionals, we have not built responsibility for our actions into the structure of our field. We do of course have scholars specializing in ethical concerns. At Tech, that includes everything from autonomous robots in warfare to the relationship between software design and misinformation on social media.

I am not simply talking about ethics, or bias, or privacy, however, but instead a larger sense that computer scientists are responsible for how their products can be used or even abused. Our philosophy must catch up to the reality of our influence.

In conclusion, I am excited by this hearing. Advances in AI are central to our economic and social future. The issues are being raised here can be addressed with thoughtful support for robust funding in basic research in artificial intelligence—including research in AI transparency and new concepts in computing theory; support for AI education throughout the pipeline; and in developing standards for the responsible use of intelligent systems. These are all areas in which the funding power of the National Science Foundation and the National Institute of Standards and Technology can make a big difference.

I thank you very much for your time and attention today. I look forward to working with you in your efforts to understand how we can best develop these technologies to create a future where we are partners with intelligent machines.

Thank you. This concludes my testimony.