Trustworthy AI Litepaper

A short version of Mozilla's Trustworthy AI Paper

Summary

What is AI?

AI stands for artificial intelligence. AI is the simulation (think copycat) of how humans think and act done by machines. The machines typically gather information (like data) each time they are used by humans to get smarter so they become more "human-like" and better at whatever it is they are designed to do. Real world examples of AI include systems that are able to understand speech or images, self-driving cars, facial recognition, online recommendations we see on Youtube or Netflix, and web searches.

What is Trustworthy AI?

Mozilla uses the term Trustworthy AI to talk about the kinds of AI that are proven to be safe enough for humans to trust. "Trustworthy AI" means that the AI:

- Is careful about the privacy of humans using it
- Is clear about the data it is collecting, and how it uses data to make recommendations
- Considers its impact on humans and their emotions, and adjusts its actions to avoid harm (and unfair manipulation)
- Can be stopped if it is causing harm to humans.

Why is it important to create AI that is trustworthy?

AI can help to improve our lives, but as it is so powerful it could also do a lot of damage. Mozilla's research has shown that the way AI is currently developed can sometimes hurt human well-being. Mozilla wants to hear from lots of different people about what they think about AI, so they can make recommendations about how to make AI trustworthy and safe to use. This will help big companies and governments to collaborate with each other to create better technology that is helpful and good.

Introduction

How does AI work now - and how could it be done differently?

AI "learns" like a kid does in school. Its teachers - the programmers - give it lessons. Instead of textbooks, they teach the AI on data. Data is information, and it could be things like names and numbers, or it could be ideas like "kids who like watching Roblox videos also like watching Minecraft videos." The AI then uses the information it has learnt to solve problems, such as: "what video do I recommend next to this person watching a Roblox video?"

This can be problematic, as we do not always know what data is being collected from us, or how it's being used. It also isn't always correct. For example, not all kids that watch Minecraft videos also like Roblox videos. And maybe, kids don't want the AI to know all about the things they like!

So what could a Trustworthy AI look like? Would kids and grown ups get to decide how it works? Could we choose what information we would like it to know (to get good recommendations)? How can we keep other information we don't want to share private?

Who is currently involved in AI?

Consumers

These are grown-ups, kids, teachers - anyone who uses AI or a service that is powered by AI (like a search engine). It's not always obvious to consumers how their information is being collected or used, which can be a bit worrying, especially when it starts to feel like the machines are listening to what we are saying or reading our emails. When consumers ask companies these questions, the answers are sometimes very long and complicated, which confuses consumers even more.

Industry

These are the people and companies that are building AI products. If you want to make your AI really good, you need a lot of data. Sometimes, industry will take shortcuts to get that data, as they want to have the best AI, but that can be harmful to the consumers, who might not know that their data is being collected.

Increasingly, industry is trying to think about human well-being when they start to build systems. What kind of things do you think they should consider?

Regulators

This is the government: and sometimes other professional groups set up to monitor what industry is doing. These people are often thinking about and making rules about how AI can be used. AI changes so quickly that a lot of governments haven't made new rules fast enough.

What are the challenges with AI?

If AI learns wrong information, or if it learns too much of one thing and not another, it can become **biased**. This means that it makes unfair decisions. It can also invade privacy, as it is simply following the instructions the programmers gave to it. It doesn't understand the consequences or our feelings.

Some of the main problems with AI are:

- **Monopoly & Centralization:** If lots of data makes the best AI, then the company that can collect the most data, makes the best AI, and makes the most money. This makes it hard for new or small companies to compete. It also means that a few companies may know a lot about us and without rules, we might not be able to do anything about our data.
- **Data privacy & rules:** As AI needs so much data, companies sometimes secretly collect data from people using the internet, to teach their AI better. Sometimes, they hide their questions in confusing and long Terms of Service agreements, which are hard to read and understand. Even still, if you could understand the Terms of Service you might not be able to do anything about them because you can't change them and you may still need to use the service because you are told to (like at school or work).
- **Bias & discrimination:** AI can only learn what it is taught, so if it is taught something wrong (that the moon is made of cheese) or if it learns a bias (that people who wear blue shirts are funnier than people who wear red shirts), it keeps making decisions based on that information, which can lead it to make bad decisions.
- **Accountability & transparency:** Because different companies are making their own AIs, they like to keep the details secret, to stop others from copying them. However, that means that no one knows how it works - or if it is following the law. This makes it difficult for anyone to make an informed decision about using the machine or service because you simply cannot get the information you might need to make a good decision for yourself.

- **Industry standards:** Sometimes, if everyone is doing the same thing, it can be hard to stop and ask questions, and act differently or to slow down to think about the consequences of their AI. Standards are set to make sure people follow along with at least some of the same rules in all of the companies in an industry.
- **Human workers:** Some AIs need to look at lots of images to learn how to identify certain objects, and humans have to find and sort all of these images. Often these aren't well-paid jobs. This means AI is being taught by limited types of people, who represent limited opinions - not everyone. This can lead to bias and inaccurate AI.
- **Safety & security:** As AI is so new, criminals are learning new ways to use it to make trouble. Because AIs are very good at doing repetitive tasks, criminals can also use them to send lots and lots of messages at the same time or use them to gather data about you to cause harm (like taking your identity or telling you fake news so you do something they want you to do).

The Path Forward

As AI is very new, and trustworthy AI is an even newer idea, it's very hard to know what the "right" way to fix these problems is! By focusing on creating underlying principles, instead of setting strict rules, we can try to teach people values that will help them to make good, trustworthy AI.

The two key principles for trustworthy AI are:

1. **AGENCY - or, human control**
   All AI should be designed with the privacy and wellbeing of humans in mind.
2. **ACCOUNTABILITY - or, company responsibility**
   All AI creators should take responsibility over how their AI acts, and any consequences.

With these two principles, Mozilla has come up with some suggestions.

1. Change industry standards to think about trustworthy AI more

We need to make sure that the people building AI know how to make sure it's trustworthy. Some ideas to make this happen include:

- Create clear Trustworthy AI rules and guidelines.
- Teach the people who are making AI about how to make it trustworthy.
- Make sure that diverse people are involved with designing, building & teaching AI.
- Encourage investors and really big companies to invest in trustworthy AI products and companies.

2. Encourage people to use trustworthy AI instead of other AI

At the end of the day, companies make products because they want people to use them. So users also need to tell companies that trustworthy AI matters to them. Some ideas to make this happen include:

- Protecting people's privacy must be seen as the most important foundation for building AI, and is used as the starting point for making new products.
- Product guides, or "trustworthy scores" (like the health labels on your food) help users choose products based on how trustworthy they are.
- Transparency about how the AI works is a feature that users ask for and companies have to include.
- Entrepreneurs come up with new trustworthy AI ideas, and investors support them with money.
- Artists, journalists and educators teach people about how AI works, how it affects them, and help them come up with ideas to make it better.
- Citizens tell companies when they aren't happy, with petitions and by asking their elected officials to act.
- Groups that care about human rights need to learn about trustworthy AI, too.

3. Make stricter rules to keep AI trustworthy

As the technology is changing so fast, sometimes the laws can't keep up. Governments need to make sure they understand trustworthy AI so they can write good and practical laws. Some ideas to make this happen include:

- People who work for the government learn how AI works.
- Existing rules that protect people are updated to include trustworthy AI, and governments are stricter on rule-breakers.
- Companies have to tell governments and regulators how their AI works.
- Governments start to invest in and use trustworthy AI.

Next Steps

This is a very big job, but it's a very important one, and it's one that we need to act on now. We need lots of different people from different backgrounds to come together and share their opinions on trustworthy AI to make sure we have a voice in how AI is made. Everyone needs to understand how AI works and why it's really important that AI is trustworthy.

Please send us or Mozilla your thoughts. Let's let industry and the regulators know how important trustworthy AI is for kids!