Explain: Algorithmic Bias

Artificial Intelligence
Winter 2022

CHICAGO BOOTH

Does AI really have “intelligence”?

- **Background knowledge:**

  Four Components of a supervised learning application:
  - Input, desired output
  - Data
  - Trainer (training algorithm needed)
  - Predictor (deployment algorithm needed)

- **Conclusion**: The only learning source of AI: data

- **Problems?**

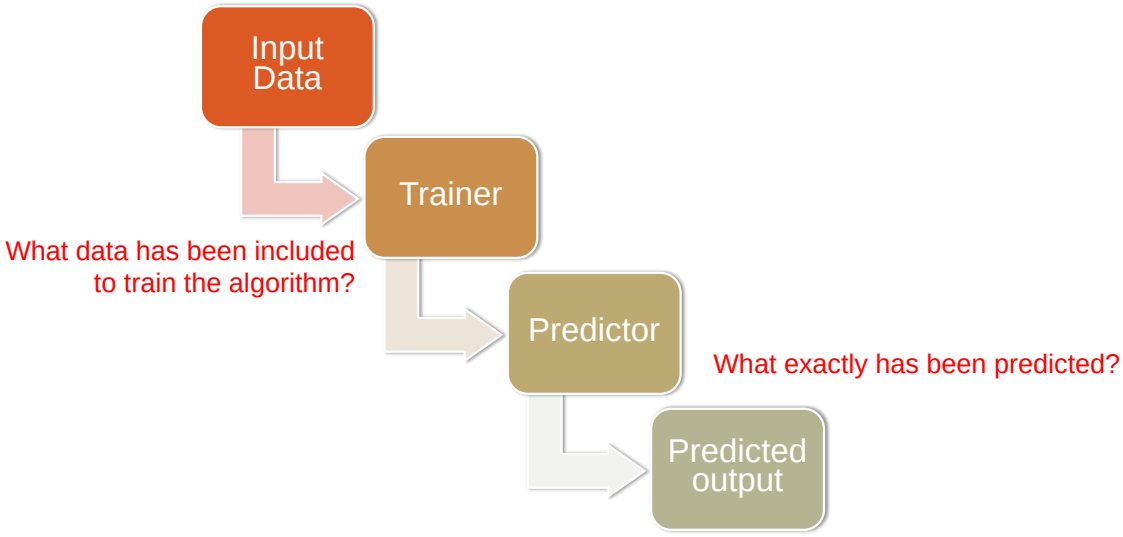
CHICAGO BOOTH

- Artificial Intelligence has a fancy name. People love to talk about it as if it has human-like intelligence.
- But today we need to clarify that AI actually doesn’t have the intelligence that people imagine it has.
- The reason will be very clear after I explain to you how exactly AI works.
- Let’s start by looking at the four components of a typical AI application under a supervised learning environment.
- First, we will need to correctly identify what is the input and the desired output. Note that the key here is to get these two things correctly.
- Then we will collect data and split it into the training and testing sets.
- Now we are at the third step! We will fit our training set into the trainer, which contains a training algorithm.
- There will be another algorithm needed for the deployment of our predictor. So, there are two algorithms needed in total for an AI task.
- This whole process I explained just now is what happens inside of the black box of AI. There is no magic!
- Again, AI does not have human-like intelligence. The only source of learning for AI is the data that people put into it.
- Wait… Do you see any potential problems here? AI doesn’t have the ability to think independently, so what will happen if there is something wrong with the

input data? Is it possible that the algorithms are not predicting what people wanted them to predict?

~ 250 words
~ 2.5 minutes

Two Problems Lead to Bias



CHICAGO BOOTH

- If we take a closer look at the process of an AI application, in other words, what exactly happens in the black box of AI, two major potential problems can be detected.
- First, what data has been included to train the algorithm? Recall the only learning source of AI: data. AI does not have human-like intelligence and is strictly data-driven.
- But oftentimes people are not aware of “what” data they put into the trainer. They may know about the size and some features of the data but lack a deeper understanding of it.
- Sometimes this problem is caused by overconfidence, arrogance, or laziness. Don’t misunderstand me, I do love humans, but you probably can’t imagine what serious results can be caused because of problematic training data.
- We will talk more about it in the next slide.
- The second problem may come from the predictor. When people are excited about accuracy scores or other measures that they think can prove the predictive power of their AI models, a much more important question would be, what exactly has been predicted?
- It is possible that your model does achieve a high accuracy score and has high predictive power, but you are not actually predicting the thing that you wanted to predict.

- It is like you aim at a target, let’s call it your label Y, but you actually shot something that is three inches from your target. The difference is so small that you may easily overlook it. In this case, you will be celebrating a “nice shot” that you did not really achieve!
- These two problems in the process of an AI application, either from the flawed training data or from the wrongly predicted label, lead to bias.
- This is the intuition behind “algorithmic bias”.

~ 300 words
~ 3 minutes

Algorithmic bias type 1:
Non-diverse training data

- The two algorithms are aimed at the right target
  - but they fail to represent underserved groups.

- Causes biases in AI

  - Example from Healthcare: A medical algorithm that was trained on primarily White patients performed poorly in Black patients.[1]

[1] *Michael W. Sjoding, et al. "Racial bias in pulse oximetry measurement." New England Journal of Medicine 383, no. 25 (2020): 2477-2478.*

CHICAGO BOOTH

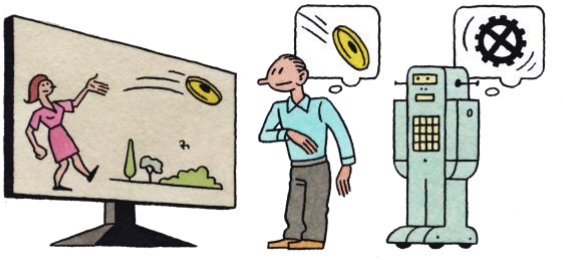
- Now, let's look at the first type of algorithmic bias. As we just discussed, this type of bias is caused by non-diverse training data.
- The two algorithms, one for the trainer and the other one for the predictor, work just fine by aiming at the right target. In other words, what the algorithms actually predict is what people wanted them to predict.
- However, these algorithms fail to represent underserved groups when deployed in reality. Why?
- Remember that we said before, the only learning source of AI is data.
- This type of algorithmic bias is caused because the training data is not representative enough, or data from some minority groups were simply not included in the beginning. How can AI predict about a group that it has never seen or did not have enough data to learn from?
- Here is an example from the healthcare industry.
- Researchers found a medical algorithm performed poorly in Black patients. Take a moment to guess the reason.
- You probably got it right. This algorithm was trained primarily on White patients, so it knows very little about non-White patients. The training data was not diverse or representative.
- Predictions made by this medical algorithm is not reliable at all because it has reinforced discrimination.

- Algorithmic bias can seriously harm the interest of minority groups and we should do our best to avoid it.
- If the training data is fine, does that mean there won't be any algorithmic bias? In fact, there is another type of algorithmic bias. We will talk about it in the next slide.

~ 260 words
~ 2.5 minutes

Algorithmic bias type 2:
ideal label VS. predicted label

- Recall the previous example from laws
  - oftentimes “proxies” cannot capture the reality
  - *Example from Laws: A prison system’s minimum height and weight requirement for hiring that discriminates against female applicants.*[1]



- A mismatch between the ideal label and predicted label

[1] *Robert P. Bartlett et al. “Algorithmic Discrimination and Input Accountability under the Civil Rights Acts (August 1, 2020).”*

Drew by Joost Swarte

CHICAGO BOOTH

- The second type of algorithmic bias comes from the predictor.
- This interesting cartoon on the right highlights that AI understands things differently than humans. Why? Because AI is data-driven and does not have any human-like intelligence.
- suppose a person tells us that a particular photo shows people playing Frisbee in the park.
- We would naturally assume that this person can answer questions like “*What is the shape of a Frisbee?” “Roughly how far can a person throw a Frisbee?” “Can a person eat a Frisbee?” “Can a three-month-old person play Frisbee?”*
- Although algorithms can label images like “people playing Frisbee in a park”, they cannot answer those questions. They simply have no idea what those questions are about.
- Again, this is because algorithms can only learn from the training data that people put into them and make predictions strictly based on the data.
- If there are differences between the ideal label that you *think* the algorithms will be predicting and the label you chose for the training data, for example, you labeled an image of “*people playing Frisbee in a park*” and wanted the algorithms to predict “*whether a person can eat a Frisbee”,* you would be distant from the truth.

- Here is another example from the laws to give you a better idea.

- In 1977, the Supreme Court ruled against a prison system's minimum height and weight requirement for hiring. This prison wanted to hire those with strength and thus used minimum height and weight as *proxies* for strength. They did not use actual strength tests.
- Eventually, the Court ruled they were discriminating against female applicants.
- This is a good example showing that oftentimes "proxies" are not enough to capture reality. It is so easy for people to make this mistake because some characteristics can be hard to measure in reality. Using "proxies" seems to be reasonable but is in fact inaccurate and deceptive.
- There is a distance between the ideal label that people wanted to predict and the actually predicted label by the algorithms.
- This mismatch between the ideal label and the predicted label will lead to algorithmic bias.

~ 350 words
~ 4 minutes

Who is easier to be fixed:
Biased humans or biased algorithms?

***Recommended Reading**:*
"Biased Algorithms Are Easier to Fix Than Biased People," Sendhil Mullainathan, New York Times, 12/6/19

- Think about this open question for a minute…
- AI systems are incredibly narrow in what they can do
- They can do certain tasks, but human-style generalizations do not apply.[1]
- Solutions to fix algorithmic bias
  - Be careful at the blueprint stage: when and how will the algorithms be used?
  - Automated human check: how far is the predicted label from the ideal label?
  - Retrain the model on a more accurate label

[1] "The Seven Deadly Sins of AI Predictions," Rodney Brooksarchive, *MIT Technology Review, 10/6/17*

CHICAGO BOOTH

- We've talked about two types of algorithmic bias and now you might be thinking, hmmm, AI is not that useful or helpful.
- It is true that AI systems are incredibly narrow in what they can do given the data-driven nature of AI.
- A person who knows how to bake apple pies very likely knows how to back strawberry pies. But for AI, they can do certain tasks but may not be competent to do others. Human-style generalizations do not apply.
- But remember that the reason why AI can be biased is simply that people who designed and deployed it are biased. It is *our* fault, not the fault of AI itself.
- Think about this open question for a minute: who, in your opinion, is easier to be fixed? Biased humans or biased algorithms?
- Social scientists have done plenty of studies on biased humans, here I want to point out a few solutions to fix biased algorithms.
- First, make sure of when and how will the algorithms be used. You need to be careful in the early blueprint stage.
- Second, conduct an automated human check on whether the predicted label is the same as the ideal label.
- The third way is to retrain the model on a more accurate label when it's necessary.

- There definitely are other ways to fix algorithmic bias and I encourage you to think about them.
- Now, let's go back to the question that you were asked earlier. I'd like to tell you the last story of today.
- In an article published in the New York Times, the author Sendhil shared his childhood experience of taking a family photo when he and his family first moved to the US.
- They expectantly opened the envelope when the family photo was mailed to them, but only to find disappointment inside. Their faces were barely visible. Only the whites of their teeth and eyes came through. They learned, much later, that the equipment had been calibrated for white skin.
- In fact, this is an experience shared by many people with darker skin. Sendhil concluded in the article, "it is much easier to fix a camera that does not register dark skin than to fix a photographer who fails to *see* dark skinned people."
- Do you agree or disagree with this conclusion, given what you've learned about algorithmic bias today?
- As I said, it's an open question and there are no right or wrong answers. I'll leave you to make that judgment and look forward to hearing your thoughts.

~ 400 words
~ 4 minutes