Artificial Intelligence Essentials: Responsible AI

Ben Weissman

Course Overview

Hi. My name is Ben Weissman, founder and data passionist at Solisyon. Welcome to my course, AI Essentials: Responsible Artificial Intelligence. Have you made your first experiences with implementing AI-powered solutions and have been wondering if there are any boundaries that you should adhere to? Did you maybe run into a situation where you used an AI-driven device or service and wondered if all the data being used and collected was being governed properly? Or did you maybe witness a situation where an AI algorithm came up with a result where you questioned whether this process was free of bias? This course will give you an overview of what responsibility in the context of artificial intelligence means and why it is important. Some of the major topics that we will cover include the meaning of responsible AI, the roles of ethics and bias to create responsible solutions using artificial intelligence, and how to implement a responsible AI strategy. By the end of this course, you will know what to look for when establishing a responsible AI strategy in your organization and how to address issues. Before watching this course, you should be familiar with what AI is and which use cases might be right for your organization so you can apply the practices from this course to them. I hope you will join me on this journey to learn more about implementing artificial intelligence the right way with the AI Essentials: Responsible AI course at Pluralsight.

Responsible and Ethical Artificial Intelligence (AI)

Responsible AI

Hello and welcome on your path to learn how to implement artificial intelligence processes in a responsible way. In this course, we will talk about what responsible artificial intelligence means, the role of bias in artificial intelligence, as well as the role of ethics in artificial intelligence and also what it takes to implement a responsible AI strategy for your organization. A quick warning before we get started. The principles and examples mentioned in this course can and will never be complete because there's just way too many use cases and angles in artificial intelligence. See them as examples, rough guidance, and ideas to then develop your own framework and come up with your own specific examples and guidelines. Let us start with the definition of responsible artificial intelligence. When we speak of responsible artificial intelligence, there are three main facets or components that we need to take into account. Bias. Is our model or solution as free of bias as it can be? Every human has bias. All AI solutions are created by humans, which inevitably leads us to artificial intelligence being biased. The question is to which degree and how are you dealing with it? And are you even aware of your specific biases so you can address them? Ethics. This is even a bit harder than the bias part because a lot of it is more about opinions than anything else. While there may be easier or more clear cases, who gets to decide if it's ethical for a companion robot to care for the elderly, a bot on a website to give relationship advice, or automated machines eliminating jobs? The last component are legal aspects, but we won't talk much about those as they are simply given. That it would be irresponsible to build a solution that is designed to commit a crime or behavior legally in another way is pretty obvious, so let's not waste any time on this. But we will talk about more in depth about the meaning of ethics and bias in AI in a few minutes. Now you may be wondering what's in it for you? These components are the pillars of a responsible AI framework and AI governance. By default, artificial intelligence is unregulated and, in many ways, often a black box for most people involved. Acting responsibly in defining your responsible AI framework is the only way to gain your stakeholders' trust, and that trust is the only way for people to approve of your solution and adopt it as long as they are given a choice.

Bias in Artificial Intelligence

One of the biggest issues we have in artificial intelligence is bias. There are mainly two types of bias, conscious bias and unconscious bias. When we talk about conscious bias, we mean that you are fully aware of your bias. And if you ask me, that means that in most cases, there should not be a place for that in artificial intelligence. Either you work around it, which is possible in most cases, or if you really can't because the algorithm is based on purpose or by design. And we will see an example where this is exactly the case, be very transparent and open about it. Let's therefore focus on unconscious bias for now and take a look at the most common cases of types of unconscious bias. Gender. Women are bad drivers. Men are horrible at taking care of children. Sounds familiar? Or the number of times at tech conferences where people assumed a woman must be in marketing because they couldn't imagine them being highly technical. These may be well-known stereotypes, and there are many more of that that we've all heard of. Despite almost all of them being proven wrong by strong evidence and data, and still, if they are in our heads, they may end up in our models. But this doesn't stop at stereotypes. If you assume that there are only two binary genders, you're entirely excluding non-binary people. You may also not have proper data about people's gender at all, leading to unfair or wrong decisions very quickly. Age. Similar to gender, we often discriminate against age. Phrases like, "We can work with them. They are too old to understand modern technology." At the same time, there are certain decisions where factors like gender or age are of relevance. So just because someone is suggesting on incorporating one of these factors does not automatically make it a biased decision. Attractiveness. If something looks good, it must be better. If someone looks good, they must be smarter, nicer, or more fun, right? Of course not. What does that have to do with AI? Think about not being invited to a job interview because some algorithm determined you're not being attractive enough. Affinity. Again, very similar to attractiveness bias. Affinity bias describes our tendency to lean towards people that we have something in common with. That may be someone who went to the same university, despite you've never met. Someone with the same hobbies. All this leading to, for example, managers leaning towards employees they can personally relate to while potentially undervaluing or overseeing fresh ideas. Attribution. Instead of evaluating a situation, person, or result objectively, we often judge them based on attributes, which can, for example, be stereotypical again, or also group-based, meaning we take characteristics of an individual and interpolate them on an entire group. Confirmation bias, meaning we tend to interpret data, experiments, or other results in a way where they match our expectations, because who wants to be wrong, right? Authority bias. We tend to believe or at least accept what our superiors tell us, or at least act based on their thoughts. But bosses are often wrong. Their views may not be beneficial for the overall success of our project, so challenge them when needed rather than to avoid an unpleasant situation. The halo effect. Kind of similar to authority, we tend to agree with people's opinions if we like the person. But friends can be wrong, too. And also, the halo effect's opposite, the horn effect, meaning we disagree with something simply because someone that we don't get along with said it, an effect that can often be witnessed in politics, where it sometimes seems like parties simply disagree with each other for the sake of disagreeing. Again, this is most certainly not a complete list, but those are the most common cases of bias, and I'm pretty sure you came across some of these already in your personal life, as well as your professional life. Okay, so we can agree that we're all biased in some way. But where and how does that bias end up in our AI solutions? In AI, we usually have a model. That model will be fed by some kind of input, and we'll be trained with a training dataset, and then generate some kind of output. The model itself is usually created by humans because we're not at the point, at least not yet, where AI solutions are being created by another artificial intelligence application or service. That means we have two sources where our bias can sit. It can sit in the data, or it can be caused by humans that created the model, which is what we call societal bias. While societal bias is usually simply driven by a lack of awareness, which, again, is mostly caused by lack of diversity on your team, and I'll also give you a few examples for that, data, or algorithmic bias, mainly originates from two issues, correlation versus causation. This means two datasets or events correlate, which can, for example, be seen in similar graphs without having to do anything with each other. A very popular example, in the 2000s, the average consumption of mozzarella cheese per capita highly correlated with the number of civil engineer doctorates awarded. While I'm sure many of them enjoy good cheese, using this as a trigger in our algorithm might be a bit far fetched. In the same time frame, other examples would be per capita consumption of chicken versus the amount of crude oil imports by the US or the divorce rate in the state of Maine versus the per capita consumption of margarine. So, don't try to make a point in making your AI bot predict that less oil will be imported by eating less chicken. Another common issue in data bias is simply incomplete or non-representative data. If your dataset only includes athletes, it may not be ideal to predict the probability of someone getting type II diabetes. If your data set only includes people with a good health insurance policy, it will most probably fail at identifying the risks to suffer from consequences of lack of medical treatment across an entire city. Non-representative data can also be caused by too much data, however, as you may, for example, include records from a time since when circumstances have changed radically. If you tried to predict which weapons would be most effective in warfare, but use data from the wars of the last 500 years, the answer might be sticks and stones. Let me walk you through a couple of examples. Let's take this picture of a young white male. There's an AI service called portraitai.app. If you upload your pictures there, those photos will be converted into Renaissance-style paintings. For our young man here, the result comes pretty close. Let's try this again. With this woman, the result is very different. Her skin is much brighter and her hair looks different, too. The reason for that probably sits in the training data. There are simply very few pictures from the Renaissance era of people that have dark skin or braids. That's one of the cases where you either have to build in bias by design because you want your output to be as close to the painting from that era as possible, or you're trying to be more inclusive, giving up the accuracy against the training data. Another example, this is a street sign of King George V street in Jerusalem. This is Isabel II bridge in Seville. "In half of a mile, turn right on Malcolm 10th Blvd." Wait, what? Malcolm 10th Blvd? No, of course, this is Malcolm X Blvd, and still many navigation and guidance systems, including Google Maps, got that wrong in the beginning because they were only trained for Roman numbers rather than for exceptions like Malcolm X. This could have been due to the lack of diversity in the backgrounds and viewpoints of the development or the review team. And another example, this time on gender. We have a male child care worker and a female professor. Let's describe that in the two sentences using their pronouns, he is a child care worker, she's a professor, and put that into Google Translate, translating it into Armenian. What you may not know is that the Armenian language is one of the few languages that doesn't have pronouns. So, what is happening if we translate our result back into English? Our formerly male childcare worker became female. Our formerly female professor became male. So, rather than using the pronoun they, which is what, in my opinion, Google could have easily done here, the algorithm is apparently biased towards genders for certain professions. All these examples show you that it doesn't take much for everyday applications to become or act biased.

Ethics in Artificial Intelligence

With the importance of being aware of you and your team's bias being made clear, let's also talk about ethics in artificial intelligence. Ethics in artificial intelligence follows a few principles. And while we can certainly argue that my list is, once again, incomplete, here are the, from my perspective, most important ones. Fairness. Is your model treating everyone fair and equal as long as they're equally qualified? Accountability and responsibility. If a decision is made by your algorithm, is there a clear owner, or does it just do what it does without anyone being in charge and taking responsibility for the actions and consequences? Transparency and explainability. If an algorithm judges about me, is it explainable what led to a certain decision or outcome? Did I not get the job because I went to a university that some other bad applicants came from or simply because I do not fulfill the objective requirements? Did I get quoted a certain interest rate because of my skin color or my credit score? Contestability. If I feel or can even prove that unfair or wrong factors went into an automated decision-making, is there a way for me to get that decision reviewed and potentially revisited and reverted? This obviously strongly correlates with a clear ownership. If nobody's in charge, nobody can overrule your case. Privacy protection and security. Are you using data that you have the owner's consent for? Are you making sure that the data used and generated is being stored in a way that it can't be accessed through others? Or did you just create a smart vacuuming bot that allows everyone on the internet to take a look at my home? Safety. Did you take all the required measures to make sure your solutions or products aren't harmful? Think of a self-driving car going rogue as one of the most obvious examples or also of a smart home that can be opened and unlocked through a camera scanning your face or phone. Again, this list may be far from complete, but we will probably agree that these are the ways of measuring how ethical or how morally acceptable a certain solution is or isn't. But why is this important? Let's take a look at a couple of more examples that are reality today. Think of so-called deep fakes. That's a thing today. You can use something as lightweight as an app on your phone to make you look like a celebrity, and it looks like the celebrity is saying what do you want them to say. Now, think of this being more than fun entertainment at a party, but someone actually trying to convince a recipient of a message or speech that it came from someone else. Who controls when or when not our smart speakers are listening to us? Which of our most private conversations get recorded, transcribed, and analyzed? Does me innocently talking at dinner about how I want to go and see France again one day trigger online ads about summer vacations in Paris? Facial recognition. Who gets to know when I was at a certain location crossing a square, and how did they acquire the images used to recognize me, and where and how long is this data stored? There are services today that you can provide a photo to, and they will or at least they will try to provide you the names of the people in the picture by analyzing profile pictures from social media. A huge company in the US started using an AI algorithm to vet job applications. But especially for tech jobs, that algorithm wasn't acting in a gender-neutral way as it was looking at data that was biased by the male dominance in the tech field, even enforcing this problem. That is hardly fair. So after a couple of tries to fix the issue properly in the algorithm, the entire solution was removed from the hiring process. If you have ever posted a picture on a social network, you probably noticed that in many cases your picture will be cropped or zoomed in. To this day on some networks when a picture has a white person and a person of color in it, the zoom will most probably be on the white person. Depending on the target group of the social network, suggested pictures of videos can be of those that show more naked skin, encouraging teenagers to pose in swimwear for their videos and profiles to be on top of that list. I could go on and on and on. Should police departments be allowed to predict where crimes are going to happen? Should drones be programmed in a way so they can make their own decisions if or if not to drop a bomb? There are obviously hundreds and thousands of more examples like this, but hopefully those show how important it is to follow ethical guidelines and act responsibly in AI.

Implementing a Responsible AI Strategy

All right, we need to avoid bias and make sure our models and processes are implemented to take the principles of ethics into account. But how do you implement such a strategy in your organization? You have AI processes and models, as well as your stakeholders. All your processes and models need a clearly defined owner. To make sure everything happens within the boundaries of ethics and the law, your entire AI landscape should continuously be audited through tools and people. Those may be internal and external. To make sure you're in line with ethical principles and to avoid bias, you can publish your models, documentation, or even open source the entire model, which serves explainability and transparency. A diverse set of both stakeholders and auditors helps you to reduce the bias and make your processes fair, and clear ownership also serves the goal of accountability and responsibility. All of those combined again makes sure that your model and strategy can be and will be trusted. One thing that you should never forget in the field of AI, stay mindful and agile. Things constantly change. You need to adapt to that, and that is fine, even if you make mistakes. Mistakes can and will be made. What matters is how you act on them.

An Example of a Successful Process Evolution

Let's close up with an example of a successful process evolution in AI. There's a service called DALL-E created by a company called OpenAI where you can basically give a description of some kind of painting or picture, and that picture will be created using artificial intelligence, like an old painting of a robot, the photo of a child, or an expressionist dog. In the beginning when you would, for example, search for a doctor and the nurse, you would often get pictures like this, a white male doctor and a white nurse. So results were biased towards race and gender. What did they do to overcome this? In the beginning, it was only an internal team that tested the algorithm with constant updates to the model. Then, they invited a couple of external users to the process, leading to more and more model updates. More and more pictures were being created with the service. But to make sure those pictures were induced in the wrong way, they came with limited user rights. More and more users were invited, again leading to more results the team could learn from and update their models on. And at some point, the limited public access was turned into full public access, and they became so confident in the algorithm that even the limited user rights were removed, and you had full ownership of all the pictures that you created using the algorithm. So when searching for a doctor and a nurse now, we may get pictures like this, for example. Still a female nurse, but also a female doctor or a person of color. When looking for a rocket scientist, we will not only get white males either. And when looking for a professor or a childcare worker, we might as well get non-gender or racially biased results out of that.

Summary

Let us summarize this course. Responsible artificial intelligence is the only way you can generate trust for your models and processes. It is perfectly fine to make honest mistakes as long as you own them, act on them, and are transparent about them. Embrace diversity amongst your stakeholders and auditors because that's the best way of reducing bias. And don't only look at what's legal, but also look at what's ethical. So in other words, what's the right thing to do with your AI processes? Thank you so much for your time. I hope this was a useful first step on your way to understand the importance of the roles of both bias and ethics play in artificial intelligence and also the importance of them being properly governed. If you have any more questions, please let me know. I'm here for a conversation.