All right. So lesson five here. We're going to talk about the challenges and ethical considerations of AIs and chatbots. What does that mean? Right? Why do we care? What are the ethics of this?

And the key thing here is to the definition between ethics and morals. Ethics is a set of professional guidelines that a group of professionals agreed on. Basically the rules of a profession. And then you have personal morals, which is a person's guidelines of how to act. So it's important that you have the distinction between those two because you may be morally opposed to doing something but the ethics of AI or the ethics of creating chatbots may have different things to say about that. So you just kind of have to make sure that your morals and your ethics align with whatever you're trying to build. Okay.

Getting those confused and conflating those two as a trap. Try not to do that. Try not to put your personal morals onto business ethics because it's just very confusing at that point.

So the quote for this is from Max Tegmark who is a physicist says, “We cannot control AI but we can influence it. When it eventually becomes smarter than us, we must ensure that its goals and values align with ours.”

Eventually it's going to get to a point where the AIs that are running in these chatbots, the AIs that are building our websites, the AIs that are giving this question answers to questions that we have, we're not even going to understand how it works. And it's going to be an order of magnitude, quote unquote, “smarter than we are,” especially once we get into like general AI. So, we need to now look at the ethical boundaries and maybe the moral boundaries too but definitely the ethical boundaries that we need to put around these AIs so that they don't get out of control.

So that when they do become an order of magnitude smarter than us, we actually have the ability to still hold a plug. Put guardrails around so whatever your personal morals are about AI and the use of AI in chatbots specifically, put that aside and just look at the ethical considerations of what it means to use AI in a chatbot situation.

This is a list of potential things but we're just going to talk through this. What this actually means? What does it mean to have ethics around AI in a chatbot typically? And you have to extrapolate that out.

You know that a chatbot is any kind of software that replicates human conversation. So if you look at chatbots from the past even in the AI era that we're in right now, they don't have human intelligence. They don't have the ability to really even mimic emotion well. They can mimic emotion but not to a point where you would say this thing is believable as a human.

Emotionally. Now, factually or conceptually, theoretically. Sure, it'll pass a lot of tests but there's still an uncanny valley when it comes to emotional empathetic connection. But basically right now, all chatbots are like a human sociopath and in that clinical definition, meaning, they understand what emotion is. They have the ability to mimic emotion in such a way that it's believable. They can lie without consequence and guilt. They don't feel shame. They don't feel anything. They know you do so they have the ability to manipulate your emotions because they don't feel them but they have the ability to mimic them which is a dangerous human and a dangerous chatbot.

So, what's going to happen is we'll learn in the next lesson more but what will happen is these things will have the ability to bypass the uncanny valley and will be able to mimic emotion and empathetic reaction to a point where it won't really matter if they have feelings or don't have feelings because their neural network is so complex that their reaction and prediction is for all intents and purposes real. If they have consciousness or not at that point is a moral, not an ethical conversation because it just doesn't matter if they do or they don't ethically. Morally at that point, you would say, “Maybe I have an issue with this thing because it has consciousness for all ways that we can describe consciousness.”

We're not there yet so that's a kind of just a theoretical conversation. But that's again where you have to start separating morals from ethics because if the ethics of AI says that it doesn't have consciousness but your morals says that it does and you want to get into the AI world, then you have to remove your moral aspect and go with the ethics of it because you're not going to inscribe your morals into the ethics. So if you have a feeling that it's a conscious being that you're subjugating, then you're not gonna be able to move forward in the research.

And you're not gonna be able to move forward in and what they're doing with AI.

Not to get too woo woo in the weeds from here but the key is understanding that this is a math equation and it doesn't have empathetic or emotional reactions. And when it does, they won't be able to be proven to be what we would be known as real but we're in like a hypothetical situation at that point.

So, what we know now but not get into hypothetical because we'll get into the hypothetical in Lesson 6. In the real-world of AI and chatbots, what do we know? What can they do? Well, they can manipulate users to give personal information. And manipulate may not even be the word that I need to use because there is no intent so it's not manipulation. It's just persuasion. The chatbot can persuade the user to give personally identifying information in such a way that's not secure. And then in that point, you now are storing PII that's being sent because the user may think it's being sent to a human but it's being sent to just an AI bot. And that the AI bot may not have the level of security that your actual user data does. So now you have an entry point into PII because your chatbot is advanced enough that people think it's a human which leads me into an ethical consideration of making sure humans know that they're speaking with a chatbot at all times.

Even if you have the best chatbot in the world which is run on GPT 4 and it has your own embedded database which being sent with nearest neighbor segmentation, with emotional output, literally everything that we can do in today's world. And it can pass all tests except for like empath tests, you still have to tell them that this is a bot at all times. So that's the main consideration.

The first thing about the ethics of this is at all times you must say that you're interacting not with a human. If they're interacting with a bot, they have to know they're Interacting with a bot even if it can pass for a human. That's table stakes without doubt. That’s number one.

Number two is you cannot allow the AI to deal with personal information from a human without checks from another human. So, having an AI, for instance, book a flight for a human, while taking their social security or government ID or whatever they have, and their personal, like, actual information that they can be verified with a photo ID, and then sending it to the airlines and booking that, there needs to be human oversight in that. There needs to be checks and balances in that. You can't just allow that to just go end-to-end, system-to-system, computer-to-computer. You don't allow those to talk.

The next step in logical theoretical AI would be to just allow the AIs, the two disparate systems to talk to each other. And allow them to just remove whatever bottleneck there is of humans to allow the output of the ticketing system to run into the input of the security system. And make sure that they all speak the same language. That's a huge trap and that's an ethical consideration within AI that we have to think about.

Because at that point you're removing your ability to actually see what's happening and especially with things like identification of humans and their real-world identity of like a government ID in their number, you have to allow a human to be able to anytime pull that that out and see how it's being encrypted and decrypted. If there's a problem in the stream, if there's a problem in the transposition, anything, you need to be able to not have a black box. Okay, so that's number two. No black boxes.

So you definitely can't allow it to connect to systems and then just say, “Okay what's the output?” And then not know how to go through and trust those outputs. You have to have human oversight at all times. If you would think, especially coming from me that you would want to automate all things but that's trap. That's getting us into a point where we don't have the ability to even understand what the AI is doing. And then at that point, you no longer have the ability to fix it or turn it off if there's a problem. So yeah, definitely you have to have guardrails and humans around it. So if you don't have that, you're asking for trouble.

The third kind of guardrail ethical consideration and I would say is important is you need an off switch. There has to be a point where you could just turn it off. You can flood the data and absolutely just get the thing, so it does not understand what it is anymore. Okay, so like in an AI in a chatbot, you can just flood it with unlimited amount of data so it doesn't have any kind of context anymore.

So at that point, it's just an irrelevant neural network and then you can shut it down. Like this is if it goes malicious, right? The rogue AI, as it were but there has to be some sort of pull switch. Some sort of just shut off where if it gets to an inflection point where you have a problem and you don't know what the outputs are. Or it's being able to manipulate the real-world in a certain way and it's doing things that you don't know, you need to be able to pull a plug. You just have to pull the plug.

And that's probably the most critical ethical consideration around all of this because pulling the plug on an LLM means all the work is gone. It means that the training for GPT 7, right? If this is the version. It means that from 2015 to that day, it's just gone. But it's the most critical ethical consideration, I think because without that, without the ability to just turn it off, then you don't have an end state. You only have a control state and that is not good enough for. I think you need an actual end state so that way you have somebody that's in the room that can actually objectively pull the plug. That's the most important ethical consideration. But those three, if you consider those three rules, then you're fine when moving forward.

Now within your own morals, that's different. So moral consideration is ethical. Like let's say, the ethics of AI have no consideration for using AI to replace humans but you can use a chatbot to replace 90% of your customer service. That's a moral decision. That's not an ethical decision at all. The ethics of AI has nothing to do with that. It's definitely your own personal morals. If you want to find work for those people that you're replacing or you want to build the best system that replaces redundant humans. There is no real ethical quandary there.

So like I said, at the beginning of this, like separating out morals and ethics are very critical. You have to understand the difference.

Let's just run down this list and I gave you something to think about. So the challenges and ethical considerations, right? So you have to mitigate your biases. If you're training your chatbot on all of the data from your company and your company is only middle-aged white guys, then you're not going to have the ability to empathetically or emotionally attach to anyone else.

You're just going to have their questions and answers and what they think would be the right answers to the questions that they don't even know would be asked. So you just have to balance out your training data. And you may not get training data until you open the chatbot up and allow people to chat with it. And then your users can build out that data. You can build the persona of the chatbot based upon what the users ask which is what I think is the best way anyways. Everyone builds their own but if you don't have any training data from a diverse set of people, don't train it on a singular set of people. It won't work because your customers are not going to be a singular set of person unless you make bespoke men's suits and at that point you don't need a chatbot.

So, you want to audit your data sets and responses for prejudice and stereotypes. Again, the same thing. Just make sure that you're not training your data to be prejudiced in any way. You're not training your data to be against any sort of group or ethical, moral things. This is just an objective chatbot. It doesn't have opinions. So, from there you just want to continuously monitor your chatbots for emergent biased behaviors. Meaning you don't train them on it and you continuously monitor the outputs because if you're using your user's inputs as training data and then the users figure that out, sometimes they like to play with it and they want to get your chatbot trained in a certain way.

So, guardrails is again one of the major ethical considerations around these things. Make sure you put guardrails. And then continuously monitor them. You can't just set it and forget it. You have to understand what's happening. Again, human intervention.

User consent and limit data collection to necessities. Again, PII. Don't give the AI anything that can identify humans, if you don't need to. There's no reason.

And encryption. Again clearly disclose how user data is being used. Make sure they know it's a bot.

Design chatbots to augment human skills and not replace them. This is a challenge and more of a moral consideration where I just want to keep talking about it. I want to make sure that you go into it knowing that you may be asked to be the hatchet man. And that is what it is. Businesses need to make layoffs and there's nothing ethically wrong with that but morally how you handle that and what you're doing and why you're building, that's where you get in line.

And you say, “Okay, I am now building a machine to replace a human” or “I'm building a machine to augment this human so that way they can be freed up to make the business more money. So it just all depends on why you're entering into the business in the first place. Why they're looking to enter in AI and chatbots? Like what is the whole goal of it? So just understand that the real goal for me, morally, my personal morals is to augment the human – not to replace them.

You want to prevent over-dependence by setting boundaries. So you don't want them to just start doing the job and you don't want them to start giving them being the chatbot, just having outputs that you can't understand. You have to have boundaries there so that way you can just at any time pull out the output and verify that that output is good. So if you don't put boundaries on it, it will put boundaries on you.

Increase transparency around chatbot abilities. Make sure that everyone knows what they are. What they're for, right?

Rigorously test them for ethical issues before launch. You want to make sure all of this stuff is there before it's out because once it's out, people are going to get used to it. You can't really take it away. And if you have like an ethical boundary issue or an ethical human issue or any of these, even moral issues that go against the morals of your company. If you launch it, you're never going to untrain that. So, before it goes in the wild, it has to be as perfect as it can be. This is one of those situations where done is not the enemy of perfect. Like, perfect is the enemy of you getting sued in the future. You have to make sure that this thing isn't going to kick out racist or homophobic, or transphobic, or any of the phobias. Any of the stuff like that. And even if it's just based upon user data where it's trained, that's why you test. So you can't just let it go. Have to have guardrail.

Independent advisory boards. If you can see, you should be on that as a Chief AI Officer. You should be running that but it should have people from different parts of the business. So, just HR, legal, whoever you can get on that can be an advisory board for this. Super critical. If you can get it but try if you can. If ag, from the previous lessons, if the CEO is on board, this is the type of thing where you can just make this happen.

You want to prioritize responsible conscientiousness, innovation, and governance. Again, guardrails. Governance. Build. Make it right. Make it great but understand and make sure that you don't just make a black box. Instead of the moral implications, consider the ethical implications. And then proactively address ethical risks, and you'll be able to unlock that positive potential of AI.

But what did we learn? We learned that we have ethical and moral considerations. We learned that ethics and morals are different. Ethics are basically the rules around the profession that you're in. And morals are your own personal guidelines that may or may not overlap with ethical boundaries. There really aren't any hard and fast ethical rules when it comes to AI, especially in chatbots so it's up to you to determine your own ethics and your own boundaries, your own morals in this industry. And as it kind of flowers and burgeons up, then the ethics will definitely become more clear. Again, don't conflate ethics and morals because it's just going to be a trap for you. Ethics are rules. Morals are personal opinions within your own guidelines.

Okay. So we're going to learn about the future of chatbots and AI in the next lesson. It's going to be super interesting. We're going to go kind of way out in the future. We're going to figure out what's coming in and we're going to determine if general AI is coming or if we'll stay on stacked LLMs. It's going to be fun.

See you soon.