



U.S. Department of Commerce

UNDERSTANDING AI RISKS AND MITIGATIONS

Training Transcript

INTRODUCTION

The coming of generative AI has been compared to the Industrial Revolution in the scale of the changes it's likely to bring. The potential is enormous, but the risks are great too. So, it's important to learn how to use AI with caution in order to reap the rewards it can bring without causing harm.

In this course, you'll learn about AI's foundational concepts and potential benefits. You'll find out how to mitigate its risks to privacy, cybersecurity, intellectual property, and trade secrets. And you'll learn how to evaluate the content it produces for accuracy and other ethical issues.

INTRODUCTION TO AI

While on its face, GenAI seems like nothing more than a text entry program, its inner workings are not as well understood. The capabilities that it provides have foundations in the very powerful and intricate programs that have trained these systems.

Objectives: Recognize the foundations of GenAI. Identify the capabilities of GenAI.

Scenario

[Trevor is seated in his office cubicle. He's on the phone with a client.]

TREVOR: Yes! *[He does a celebratory fist pump.]* That's great. We can have contracts out to you by the end of the week. No, thank *you*. Okay. Bye-bye. *[He ends the phone call.]* Yeah! What? *[He looks at his computer screen.]* Are you reading this?

[Clinton, who works in the neighboring cubicle, pops his head over the wall.]

CLINTON: The AI Challenge? Yeah, man. I think it's gonna be fun.

TREVOR: Fun?

CLINTON: Yeah, AI. There's so many possibilities.

TREVOR: You *would* love this.

CLINTON: You *would* complain about this.

TREVOR: You're the one who's into tech. Can't even keep up with these platforms as they are.

CLINTON: What if AI could make your job easier?

TREVOR: How's it gonna do that?

CLINTON: Lots of ways. Use it to help write your proposals. You always say how long they take.

TREVOR: AI can do that?

CLINTON: Sure, man.

TREVOR: How?

CLINTON: Try reading past the first paragraph of that e-mail.

TREVOR: Ha. Ha.

CLINTON: I'll send you some articles. It'll be great, you'll see.

TREVOR: Alright, buddy.

[Clinton leaves his cubicle and enters the neighboring office, where he sees Olivia.]

CLINTON: Hey! Did you get that e-mail about the AI challenge?

OLIVIA: I did. I've read all about how it can create blogs and social-media posts. I've been dying to try it.

CLINTON: I knew you'd be into this.

OLIVIA: Oh sure! Who wouldn't be?

CLINTON: Right?

OLIVIA: How are *you* gonna use it?

CLINTON: I mean, there's so many options. I just can't decide. It can help create logos, create product descriptions, even design market-research tests. I mean, there's possibilities. They're endless.

OLIVIA: Hey. Well, just try to narrow it down in time to actually take *part* in the challenge.

CLINTON: I'll try my best.

OLIVIA: Sorry, but I have a meeting about to start.

CLINTON: Yeah, yeah. No worries. *[Clinton returns to his office.]*

OLIVIA: Hey! *[Olivia sees Isabella and gets her attention.]* Are you up for this AI Challenge?

ISABELLA: Actually, it was my idea.

OLIVIA: It was? That's great.

ISABELLA: Yeah, I'm excited to see what everyone comes up with.

OLIVIA: Me too.

Introduction to AI (continued)

It seems like everyone in the world these days is talking about generative artificial intelligence, or GenAI. Everyone has some idea of what it does – but fewer people can explain what it actually is.

Artificial intelligence has been around for a while; the first chatbots, like ELIZA, were created in the 1960s. But where ELIZA's responses were simple text reflections of user input, generative AI has created programs that can draw on vast amounts of information to create brand-new images, text, video, and audio content.

Generative AI Concepts

So how does it work? Generative AI is based on some foundational concepts.

Machine Learning

Artificial intelligence refers to the practice of getting computers to carry out tasks in a way that mimics human intelligence, and machine learning is a subset of artificial intelligence. It's a process that enables AI to learn from data patterns without human intervention. This allows AI to be far more sophisticated than old-fashioned programs like ELIZA.

Machine Learning Models

To create generative AI, data scientists and machine learning experts build machine learning models, that is, mathematical models that have been trained to identify patterns in data and then create content based on that data in response to user prompts.

Algorithm

A type of repetitive mathematical instruction that can identify patterns in huge pools of training data. The patterns identified by these sophisticated algorithms can often be invisible or even counterintuitive to human thinking.

Training Data & Machine Learning Bias

In creating the model, data scientists have to carefully screen for machine learning bias, which can happen when the training data includes biases, or when basic assumptions used to create the algorithm are biased.

Large Language Models

Generative AI began to break into public consciousness with computer-generated images, such as pictures showing what a user would look like if painted by a Renaissance artist. These days, it's large language models that are making the biggest splash. Beginning in 2010, AI researchers discovered that training models using vast quantities of text had better results than training models based on grammatical rules. By using this approach, scientists managed to create models that could make sense of words by analyzing their surrounding context. This breakthrough paved the way for today's large language models, such as the one used by ChatGPT, which can provide original text on a massive range of topics in response to user prompts.

Chatbots

Chatbots simulate natural conversation with human users and have existed for a long time. ChatGPT and other GenAI models can go well beyond simple applications such as customer service, and have already been used to create college-level essays, poems, and even functional computer code. Many chatbots are available for public use today.

Generative AI Image Creation

Other types of GenAI work with images. DALL-E, for example, can create images in innumerable styles based on prompts from a user; if you ever wanted to see, say, the moon landing as painted by Salvador Dali, DALL-E can help you out.

GenAI models use deep learning to go through raw training data, be it vast swathes of text on the internet or the collected works of Salvador Dali and other surrealist painters, and use what they find to generate statistically probable outputs.

That's how they can create new content that is similar, but not identical, to the original data.

The Growing Potential of AI

The key to these models' success is their ability to learn by themselves, meaning researchers don't have to label every piece of data that the model is given – the model is able to classify data by itself. That means it can receive more data than it could if researchers were limited by their own constraints in labeling data, which in turn means the model can grow and increase its capabilities more quickly. The potential is truly extraordinary.

Create New, Original Content

GenAI can be used to create new, original content like images, videos, and text, which businesses can use, for example, to help generate new ideas for marketing content or product development.

Automation and Acceleration

It can be used to automate and accelerate tasks and processes, saving time and money in all sorts of industries and business processes. For example, companies are already using GenAI to generate marketing content.

Explore and Analyze Complex Data

GenAI can also be used to explore and analyze complex data in newer ways. In medical research, for instance, generative models have been used to develop new proteins to accelerate drug discovery. And in climate sciences, generative models are helping to create climate models and predict natural disasters.

Improve Efficiency and Accuracy of Current AI Systems

GenAI can be used to improve the efficiency and accuracy of existing AI systems by creating synthetic data that can be used to train and evaluate other models. For example, in the automotive industry, synthetic data is being used to help develop autonomous vehicles. Road-testing self-driving automobile systems in virtually generated worlds decreases risk while also enabling testers to make changes faster and at lower cost.

Considerations for Generative AI

GenAI can do a lot, but it's important to remember that the intelligence is artificial – the machine may seem to be thinking, but it's not. GenAI is best used as an assistant, not as a fully competent professional, and all its outputs need to be reviewed by someone who is an expert in the field.

As such, **there are two important considerations you must account for:**

- **Monitoring:** You're responsible for both the inputs and the outputs of a GenAI program. So be sure that you've thought hard about the input you give it, and that you've fully vetted anything produced before you publish or share it. Don't expose yourself or your business to negative outcomes if the outputs prove to be inaccurate or inappropriate.
- **Safeguards:** Whatever the risks, the benefits are so great that GenAI looks sure to be a feature of life for a long time to come. Thus, it's essential to put in place

safeguards against misuse, and to hold people accountable if they fail to observe these safeguards.

Knowledge Check: Introduction to AI

1. In the course, you were introduced to some concepts that are foundational to generative AI. Select the concepts that are necessary for a complete understanding of generative AI.

- A. Large language model
- B. Chatbot
- C. Machine learning model
- D. Algorithm
- E. Training data
- F. Machine learning bias
- G. Personally identifiable information
- H. Data scientists
- I. ELIZA

Answer: A, B, C, D, E, F

Explanation:

- A. This option is correct. Large language models provide original text based on user input on a massive range of topics.
- B. This option is correct. Chatbots simulate natural conversation with human users.
- C. This option is correct. Machine learning models are mathematical models that have been trained to identify patterns in data and then make predictions, recommendations, or decisions based on that data.

- D. This option is correct. Algorithms are a type of repetitive mathematical instruction that can identify patterns in huge pools of data.
 - E. This option is correct. Training data is used to train AI models.
 - F. This option is correct. Machine learning bias can happen when training data includes biases, or when basic assumptions used to create the algorithm are biased.
 - G. This option is incorrect. Personally identifiable information is **not** a foundational concept of AI. However, it is important to ensure that personally identifiable information is not exposed to AI.
 - H. This option is incorrect. Data scientists work with and develop AI, but they are **not** a foundational concept of AI.
 - I. This option is incorrect. ELIZA was the name of an early chatbot; it is **not** a foundational concept in AI.
2. GenAI presents many opportunities and benefits. Select the potential benefits of using GenAI.
- A. It can create new, original content like images, videos, and text.
 - B. It can improve the efficiency and accuracy of existing AI systems.
 - C. It can be used to explore and analyze complex data in newer ways.
 - D. It can automate and accelerate tasks and processes.
 - E. It can remove the need for fact-checkers.
 - F. It enables companies to use other people's intellectual property.
 - G. It can substitute for human judgement in all cases.

Answer: A, B, C, D

Explanation:

- A. This option is correct. GenAI can be used to create new, original content, including images, videos, and text which businesses can use, for example, to help generate new ideas for marketing content or product development.
- B. This option is correct. GenAI can create synthetic data that can be used to train and evaluate other models.
- C. This option is correct. In medical research, for example, generative models have been used to develop new proteins to accelerate drug discovery.
- D. This option is correct. Automation of tasks and processes can save time and money in all sorts of industries and business processes.
- E. This option is incorrect. All of the content that AI produces should be checked by a human, to ensure accuracy.
- F. This option is incorrect. One danger of using AI is that it may violate intellectual property rights. For this reason, it's important to review AI content.
- G. This option is incorrect. AI is an assistant, not a full-fledged professional, and it should always be overseen by a competent human.

PRIVACY AND CYBERSECURITY RISKS

In the race to make better GenAI tools, sometimes shortcuts are taken by developers, leaving potential security vulnerabilities. On top of that, people who look to use any and all of these tools sometimes put too much faith in these programs and can run the risk of getting themselves or their organizations into serious trouble if they don't know how to protect themselves.

Objective: Recognize the risks to privacy and cybersecurity when working with GenAI tools.

Scenario

[Trevor is walking into the office, fixated on his cellphone. He doesn't notice Clinton walking toward him with several small boxes. Clinton bumps into him and drops one of the boxes.]

CLINTON: Dude! What are you doing?

TREVOR: Oh, man. Sorry! *[Clinton bends down to pick up the box.]* But hey, guess who landed his third high-dollar deal this month? This guy! *[Trevor points to himself.]*

CLINTON: Great job. Go you.

TREVOR: Thank you.

CLINTON: You say your third one this month?

TREVOR: Yeah.

CLINTON: How did you manage that?

TREVOR: Well, I'm using the AI on my proposals, just like you suggested. I think I'm gonna buy a boat with all the commissions I'm about to rake in.

CLINTON: Haha! I told you, man. I bet you like AI now.

TREVOR: Like it? I love it!

CLINTON: Haha! Well, maybe next time you will listen to me.

TREVOR: Eh, there's always a chance.

CLINTON: Ha! Hey, did those articles I sent you help?

TREVOR: C'mon, who reads the manual? No, I played around with the prompt engineering. It didn't do exactly what I wanted so... I improvised.

CLINTON: Haha! Man, look at you. You almost sound like you know what you're talking about.

TREVOR: Laugh it up. No-one's gonna rain on this boat parade.

CLINTON: Hahaha. Okay, so you refined your prompt engineering. What else did you do?

TREVOR: I found this article explaining how to input data from the sales tool... *[Isabella enters the office and pauses as she hears this.]* Took it to a whole new level.

ISABELLA: Importing data from the sales tool? I didn't realize internal AI allowed for that yet.

TREVOR: Well, technically it doesn't.

ISABELLA: Please don't tell me that you imported internal sales data into a random AI that's available to the general public?

TREVOR: Just a little. But – I got a sale.

ISABELLA: A little?

TREVOR: Yeah, for my proposals, but that's it. And it works, so you can't argue with results. Am I right?

ISABELLA: Trevor, you can't do that. That data can be accessed by anyone now.

TREVOR: Anyone?

ISABELLA: Different platforms have different privacy settings. Did you even read the terms of service?

CLINTON: Seriously? *[Clinton shakes his head in disbelief and walks away.]*

Privacy and Cybersecurity Risks (continued)

Users of GenAI tools like ChatGPT have found all sorts of ways to increase their productivity and make quick headway with routine and nonroutine tasks.

For example:

- You might use ChatGPT to convert facts and figures into a handy slide deck for a big presentation.
- You might use it to help you refine your approach to a marketing strategy.

But what happens if those facts and figures contain confidential business information, or if your marketing strategy is based on personal information obtained from your customers?

One key mistake that people make in using GenAI is assuming that the information they share is secure and safe, and that sharing data with GenAI is not violating anyone's privacy rights. But that's not necessarily the case.

Potential Risks

When you work with GenAI, it's essential to remain aware of the privacy and cybersecurity risks, so that you can get the most out of these tools without exposing yourself or others to harm.

Poor Development Process

Companies are deploying GenAI tools extremely rapidly, which means that the usual controls for ensuring appropriate development and management might not be in place. Companies may not have set up proper protections to ensure privacy, security, or anonymity.

Data Breaches and Identity Theft

GenAI tools may be vulnerable to cyberattacks. And yet, they require a great deal of information from the user to operate. That could open the door to data breaches and identity theft, if the data the user provides is exposed to bad actors.

Poor Information Security

Integrating GenAI tools into your systems can be a complicated business. The algorithms GenAI tools use are extremely complex, and they iterate constantly, which makes it hard to screen for vulnerabilities that may not even exist yet. This could mean poor security for your systems as a whole. Malicious actors may be able to trick the AI into classifying dangerous input as safe, which could provide a back door to get malware into any system that integrates the AI.

Data Leaks

Remember that slide deck for the presentation? If, say, your presentation was a pitch for future business, you might provide the AI with confidential information about your potential client's needs and vulnerabilities. If a competitor were to later ask the AI for information on your client's vulnerabilities, the AI could potentially use the sensitive information you gave it to answer that question.

And it's not a hypothetical; employees have shared trade secrets, classified information, intellectual property, and personally identifiable information about customers with GenAI tools.

Along with the business risks, just the act of sharing that personally identifiable information with the tool could expose your business to liability under laws like the European General Data Protection Regulation.

Malicious Use of Deepfakes

Another cybersecurity risk presented by GenAI is the malicious use of deepfakes. With voice and facial recognition becoming more and more standard as access control measures, the increasing ability of GenAI to copy likenesses and voices may pose serious security issues in the future.

AI Governance

To mitigate these risks, businesses need to define an effective AI governance strategy that lays out principles and best practices for the use of AI, as well as monitoring frameworks to make sure everyone lives up to the standards set.

Businesses should consider:

- Whether AI use is necessary.
- What data is appropriate to use in an AI model.
- Whether the model is performing as expected.
- How best to follow compliance and reporting procedures.
- Whether bias is being introduced.
- Whether privacy is being respected in inputs and outputs.

Setting up guardrails like this can ensure that AI is used safely and appropriately to help your company reap the benefits, and also help to prevent sharing of information that might be considered private, sensitive, confidential, or personally identifiable.

If you're not sure, don't share it. The benefits aren't worth the potential damage.

Knowledge Check: Privacy and Cybersecurity Risks

1. A doctor uses GenAI to track a patient's medical history, maintain images of their facial features and identifying marks, write letters recommending treatment, and automate billing to their insurance. Select the privacy and cybersecurity risks involved in this scenario.
- A. Poor development processes in AI could expose the patient's data.
 - B. The patient could be at risk of identity theft.
 - C. Without proper data security, malware could be introduced to the doctor's systems.
 - D. Without proper data security, the patient's medical history could be published on the internet.
 - E. Patient privacy is compromised through the use of deepfakes of patient medical images and records created for training purposes.
 - F. The insurance company could sue the doctor for fraud.
 - G. The doctor could recommend the wrong treatment.

Answer: A, B, C, D, E

Explanation:

- A. This option is correct. Companies are deploying GenAI tools extremely rapidly, which means that the usual controls for ensuring appropriate development and management might not be in place. Developers of large language models may not have set up proper protections to ensure privacy or anonymity.
- B. This option is correct. Bad actors can find ways of accessing data through GenAI, and could use the patient's personally identifiable information to steal their identity.

- C. This option is correct. Using GenAI for so many purposes involves giving it a great deal of access to the doctor's systems, which bad actors could exploit.
- D. This option is correct. Data leaks can happen when GenAI is used with personally identifiable information that is not securely managed.
- E. This option is correct. The doctor has collected the visual information needed to create a deepfake of the patient.
- F. This option is incorrect. Based on the facts provided, there is no reason to believe that the information provided to the insurance company is incorrect or fraudulent.
- G. This option is incorrect. If the doctor was relying solely on the AI to decide on treatment, then the AI could choose the wrong treatment. But in this case, the doctor is only using the AI to write letters, not to choose treatment.

COPYRIGHT INFRINGEMENT & DISCLOSURE OF TRADE SECRETS

When using any AI tool, it's critical to remember that one must not share anything sensitive with the tool. Equally important is the need to respect intellectual property rights by never using third-party content without the appropriate rights or licenses.

Objective: Recognize best practices to follow to avoid copyright infringement and disclosure of trade secrets.

Scenario

[Clinton gets up and shows a custom-designed t-shirt to Trevor, who's seated in the neighboring cubicle.]

CLINTON: Ta-dah!

TREVOR: Wow! That's great. *[Trevor gets up to have a closer look at the design.]*

CLINTON: Everything I said it would be, right?

TREVOR: It is.

CLINTON: It's bright.

TREVOR: Yes.

CLINTON: Colorful.

TREVOR: Ah-huh.

CLINTON: And there's nature!

TREVOR: Definitely! You know what? I love it and Nichols is gonna love it.

CLINTON: Yes! *[The coworkers share a high-five.]*

TREVOR: Question, though. You usually don't do designs, so why are you in that area?

CLINTON: Two words: AI... Challenge!

TREVOR: What!

CLINTON: Yeah. And plus, we're on a deadline so I figured it might work. Turns out, I was right.

TREVOR: It makes all the sense in the world.

CLINTON: Yeah. I just put in what we wanted, I tweaked the parameters, and poof. Perfection, right? Technology's great.

TREVOR: It's that easy.

CLINTON: Well, that easy for me. For you, not so much.

TREVOR: Ha. Ha.

[Isabella enters the office. She's holding a tablet.]

ISABELLA: Hi, guys. *[The design catches her attention.]* Oh, a Cabricci! Who's the fan? I didn't think either of you had that much taste.

CLINTON: What?

ISABELLA: Cabricci, it's classic Cabricci.

TREVOR: I'm sorry. What's a... Cabricci?

ISABELLA: The artist who designed that.

CLINTON: No, no. I designed this.

ISABELLA: And you don't know who Cabricci is?

TREVOR: *Clint* designed it with AI... For the Challenge.

ISABELLA: Uh, no. That's Cabricci. It's famous.

CLINTON: It's not that famous, huh? Hahaha.

[Isabella shows them her tablet's wallpaper, which bears a striking resemblance to Clinton's AI-generated design.]

TREVOR: Oh.

ISABELLA: Yeah.

TREVOR: Wow. *[Trevor returns to his cubicle.]*

ISABELLA: Yikes. *[Isabella walks away.]*

CLINTON: It wasn't that close, right?

TREVOR: It's the exact same, bro.

Copyright Infringement & Disclosure of Trade Secrets (continued)

The potential uses for AI are increasing every day, as is the value that businesses perceive it to have. But many companies are wrestling with the tricky problem of how to balance the risks against the benefits.

Concerns about intellectual property are giving some business leaders sleepless nights over both the risk of copyright infringement, and the risk of providing AI tools with intellectual property that could be made public, damaging business strategy and potentially undermining trade secret protections.

Generative AI tools are trained on vast quantities of data that are sometimes obtained in unclear ways. Algorithms use these datasets to generate content that is similar to, but not the same as, the original data provided.

Best Practices

What happens when the training data was somebody else's intellectual property?

That's a question that hasn't yet been answered, and that courts are currently considering. Currently, a group of artists are suing multiple generative AI companies, saying that the platforms used their work without consent to train AI to make art in their styles. They contend that the work produced is not transformative from their original work, and therefore is unauthorized, derivative work.

Other cases have been brought by authors who say that AI tools have been trained using their copyrighted writing without their consent. More cases are likely to follow, which means there's a great deal of uncertainty around the use of work generated by AI.

Avoid Using Someone Else's IP as Your Own

For users of these tools, the important thing to remember is to avoid using someone else's intellectual property as your own, so as to avoid plagiarism or copyright violation. Always check the outputs of AI to make sure that the work created is not identical, or overly similar, to someone else's existing work.

Avoid Submitting Your Own or Your Organization's IP

Generative AI tools are often trained on user inputs, which means that once you've entered your intellectual property, it could be used to respond to other people's inputs. This could mean the damaging disclosure of business information that should be kept private. Say, for example, you're having a disappointing quarter with regard to one particular product line, so you use a chatbot to help you write a strategy document as to how to get things back on track.

There are circumstances under which, by using the correct prompt, another user could, either on purpose or accidentally, cause that chatbot to provide it with your confidential sales information. And if the information given to the AI includes personally identifiable information, such as customer names or employee salaries, you could be exposed to serious legal hazards. To make sure this doesn't happen, avoid submitting your or your organization's intellectual property or other private information to any GenAI tool.

Avoid Sharing an Organization's Trade Secrets

Perhaps an even greater issue than an improper disclosure of data is the problem of trade secret protection. Misappropriating trade secrets can result in severe penalties, but only if the holder has taken appropriate steps to maintain the secrecy of the trade secret. Businesses go to great lengths to ensure their secrets remain secret, such as limiting the number of people with access to the secret, including confidentiality obligations in contracts, and setting up technological barriers to access. But none of those protections may matter if that secret is disclosed to a GenAI.

Say, for example, an employee submits detailed product specs to a GenAI tool as a prompt to create a response to a request for proposal. Even if no other user gains access to those specs, just the act of entering the secrets as an input to the AI could undermine all other efforts to maintain their secrecy. So, it's essential that all employees avoid sharing an organization's trade secrets with GenAI.

Ways to Avoid Pitfalls

Companies can help avoid these pitfalls by implementing training, making sure that all employees understand what – and what not – to share with GenAI. And strict data controls can help to ensure that only those who need to can access sensitive information. Some companies such as Amazon and Verizon have made rules restricting the use of GenAI, or even banning it entirely.

Most importantly, you can ensure that before you use, publish, or share content created by a GenAI system, you ask yourself, "Is the output of this content clearly attributable to some other content creator?"

And, before you share any information with a GenAI tool, ask yourself, "Is the data I am providing sensitive or confidential to the people or organization I am working with?" If the answer to either question is yes, don't share it.

Knowledge Check: Copyrights & Trade Secrets

1. Some best practices can help you avoid copyright infringement and disclosure of trade secrets in using GenAI. Select the best practices that apply to avoiding copyright infringement and disclosure of trade secrets.
 - A. Avoid submitting your own or your organization's intellectual property.
 - B. Avoid using outputs that could be interpreted as someone else's intellectual property as your own.
 - C. Avoid sharing an organization's trade secrets.

- D. Avoid providing text inputs to GenAI.
- E. Avoid using images produced by GenAI.
- F. Have a lawyer evaluate all GenAI content before using it.

Answer: A, B, C

Explanation:

- A. This option is correct. To avoid data leaks, it's important not to provide AI with your intellectual property or that of your company.
- B. This option is correct. Using content that could be interpreted as other people's intellectual property may be a copyright violation and should be avoided.
- C. This option is correct. Providing trade secrets to GenAI could invalidate trade secret protections.
- D. This option is incorrect. Not all text inputs to GenAI are inappropriate.
- E. This option is incorrect. If done responsibly, using images produced by GenAI may be valid.
- F. This option is incorrect. It's important for content to be evaluated, but it doesn't necessarily have to be done by a lawyer.

RECENCY, DELUSIONS, AND ACCURACY OF INFORMATION

For its many strengths, GenAI also has its weaknesses, in that it can be used for nefarious purposes. Many who wish to can make it say and display anything imaginable, and this causes many headaches for all who are exposed to it.

Objective: Identify the questions to ask when evaluating content generated by GenAI.

Scenario

[A social-media alert sounds as Olivia takes a seat in her office chair. She begins typing on her computer, but several additional alerts force her to look at her cellphone. Puzzled

by the numerous notifications, she drops her phone and types frantically on her computer.]

OLIVIA: What? I don't understand.

[Isabella enters the office.]

ISABELLA: What is going on with social media?

OLIVIA: It's a disaster. It's blowing up. And they're all saying we put out misinformation?

ISABELLA: Misinformation?

OLIVIA: They're saying we cited a false study claiming improvements in durability?

ISABELLA: Where?

OLIVIA: Looks like it's Serge's last blog announcing the new line.

ISABELLA: Oh. Serge has been using GenAI to create blog content, right? I hope he's made sure to check his facts.

OLIVIA: I reminded him how important this was, but he's been cranking out so much content lately. He must've missed something.

ISABELLA: Sounds like an AI hallucination.

OLIVIA: A what?

ISABELLA: Hallucination is where AI can fabricate facts and present them in a way that appears completely believable. GenAI is a great tool, but you always need to fact check.

[Another social-media alert sounds, prompting Olivia to check her computer screen.]

OLIVIA: Oh geez. Someone just called it "Blog-gate" and is calling for an investigation. What do I do?

ISABELLA: First, get the blog pulled down. And then, I'll help you draft a retraction.

OLIVIA: Thanks.

Recency, Delusions, and Accuracy of Information (continued)

Generative AI tools have become so advanced that their responses can be indistinguishable from those of a human being. But the fact remains that these are advanced computer programs; they're not actually able to think.

This is one reason why information accuracy remains a real problem in the use of GenAI. GenAI tools are designed to produce information that sounds true, but the tools are not self-policing enough to check whether it actually is true.

Is the Information True, Accurate, and Current?

The only way to address this problem is to be very careful with your use of AI outputs. When you're evaluating content produced by GenAI, always ask yourself, "Is this information true, accurate, and current?" Consider whether you have the necessary expertise to make that call; if you don't, then find an expert who does.

Can this Information be Verified by Other Reliable Sources?

When large language models generate false information, it's referred to as a "hallucination." These hallucinations sound plausible, because that's what these models are designed to do: create fluent, coherent text that makes sense as an answer to a prompt. But "makes sense" is not the same thing as "true," and using false responses can land you in serious trouble, as one lawyer found when he filed a brief based on arguments made in cases that didn't exist, but had been made up by the AI. In other cases, GenAI tools have provided citations to made-up papers by real authors, on topics

that they easily could have written on, but in fact didn't. To avoid falling into this trap, when evaluating an AI response, it's important to ask yourself, "Can this information be verified by other reliable sources?" Always verify the information AI gives you. And again, if you don't have the knowledge to do so, then find someone who does, rather than relying on the assertion of an AI system.

Is There Misleading Information, Misinformation, and/or Disinformation in the Generated Content?

GenAI can also be used to create misinformation and disinformation, which can be used to mislead the consumers of this content. Already, AI-generated deepfake videos have been shared that claim to show political candidates in situations that never really happened, as well as prominent actors advertising products that they would never really endorse. It's important when you're creating AI content to ensure you don't, even inadvertently, perpetuate this kind of deception. Always ask yourself, "Is there misleading information, misinformation, or disinformation in the AI-generated content?"

Due Diligence

If you choose to use anything generated by GenAI, do your due diligence. Never assume that the outputs of AI are true, accurate, or current without first verifying that they are, either through your own expertise or someone else's.

Knowledge Check: Recency, Delusions, & Accuracy

1. Asking questions can help you properly evaluate the information generated by AI. Select the questions that can help evaluate AI content.
 - A. Is this information true, accurate, and current?
 - B. Can this information be verified by other reliable sources?
 - C. Is there misleading information, misinformation, or disinformation in the generated content?

- D. Does the information sound plausible to me?
- E. Do I understand exactly how the AI generated this response?
- F. Does the content cite authors who I know to exist?

Answer: A, B, C

Explanation:

- A. This option is correct. Consider whether you have the necessary expertise to make that call; if you don't, find an expert who does.
- B. This option is correct. Large language models are prone to hallucination, which means creating plausible-sounding responses that aren't in fact true. It's important to verify all content they produce for this in order to address this problem.
- C. This option is correct. GenAI can also be used to create misleading information, misinformation, and disinformation, which can be used to mislead or defraud the consumers of this content. Make sure your content doesn't perpetuate this problem.
- D. This option is incorrect. GenAI is designed to sound plausible, so merely sounding like it's true doesn't mean that it is.
- E. This option is incorrect. It's not necessary to understand the technical specifications of how it works in order to use it safely and effectively.
- F. This option is incorrect. It's not enough to know that the authors cited by AI exist; sometimes, AI attributes fake information to real sources.

ETHICAL RISKS OF AI

GenAI was trained by all existing information that can be found on the internet. As a result, it got all the bad along with the good. GenAI can rely upon stereotypes, be

improperly biased, and misrepresent groups and individuals. Knowing how this happens can help you avoid using, displaying, and sharing this incorrect information.

Objective: Recognize the ethical considerations when utilizing GenAI.

Scenario

[Simone is seated at her office desk. She sighs as she examines some forms. Then, Isabella enters the office.]

ISABELLA: Something wrong?

SIMONE: Yeah, these. HR gave me what's supposed to be the best candidates for the director position, but I don't get it. I know some great people who should be in this group and they're not here.

ISABELLA: What do you mean?

SIMONE: There's several people who I know are more than qualified and they're not here.

ISABELLA: And you know that they applied?

SIMONE: Yeah. Ramon came to see me the day after the posting went up and he told me he'd already applied.

ISABELLA: And he's not in there?

SIMONE: No, and Yvette isn't either.

ISABELLA: Well, I know that she applied. We just talked about it last week. She's a stellar candidate.

SIMONE: I know. I asked her to apply in the first place. She should be here. And there are others missing too.

ISABELLA: And you're sure HR gave you the right resumes?

SIMONE: I had a whole conversation with Sonya when I picked these up. She said these are the prescreened candidates we have to choose from for the position. *[Simone passes the stack of resumes to Isabella.]*

ISABELLA: How were they screened? *[Isabella briefly examines a few of the resumes.]*

SIMONE: I don't know.

ISABELLA: These candidates are fine, but they're less qualified than Ramon and Yvette. And Sonya screened them?

SIMONE: She just said they're prescreened, not by who.

ISABELLA: We are testing out a new AI system to help screen applicants for open positions.

SIMONE: Why? The system we have works fine.

ISABELLA: Until it doesn't. It's a way of taking out the human bias in the equation. Plus, it takes our team a really long time to sift through all those applications by hand.

SIMONE: I guess that makes sense, but if this is the result...

ISABELLA: Something must be wrong with the AI system, especially if it's not giving us our best applications. Maybe they have something in common?

SIMONE: Ramon is a top performer. I don't know what could knock him out.

ISABELLA: Maybe it's an HR issue we don't know about?

SIMONE: I can't imagine what. He's such a nice guy. And I definitely can't imagine Yvette having any type of HR-related issue either.

ISABELLA: True. Yvette did take a lot of sick leave last year.

SIMONE: That's not relevant. That was protected medical leave.

ISABELLA: Maybe it's social media. I hear that that can have an effect on recruiting AI algorithms.

SIMONE: Yvette doesn't even *have* social media.

ISABELLA: Duh. Haha, how could I forget? She loves telling us she has so much better things to do. Okay. Well, I'm back to sick leave, then.

SIMONE: Ramon did just take parental leave when the twins were born.

ISABELLA: That's right, he did. Well, maybe that's it?

SIMONE: But other people were out a lot last year. And again, it's protected leave just like Yvette's.

ISABELLA: Let's send it back to the HR team to determine if some bias in the dataset or algorithm caused the model to give bad recommendations.

SIMONE: Yeah. I'll take this back to HR and find out.

[Isabella exits the office.]

Ethical Risks of AI

Generative AI is a tool, and as with any tool, the person who uses it is responsible for what it does. Computers can be trained to do many things, but human beings will always retain the duty to make ethical and legal decisions. For that reason, it's important to be aware of the ethical considerations involved when using generative AI: ensuring accuracy and avoiding illegal bias.

Machine learning models are trained on datasets that are usually drawn from the real world, which means they can often import the biases in the world around us into their thinking, and generate content that may result in decision-making on an illegal basis. The output they produce may rely on illegal and improper stereotypes about legally protected classes, potentially resulting in discrimination.

Example of Ethical Risks

For example, a study of the image-generation tool Stable Diffusion found that the model routinely created images of white men when asked to provide a picture of a CEO. Men with dark skin were portrayed as criminals, and women with dark skin were portrayed as holding low-paying jobs. Women were rarely shown to be doctors or other high-paying professionals.

The tool was trained on real-world images, but the output it produced was shown to be even worse than the training data in terms of the number of women and minorities in high-net-worth occupations. And since models like this often train on the data they create themselves, there is potential for the problem to worsen over time.

Questions to Ask

For end users of tools like this, some best practices can help ensure that bias is screened out.

When generating AI content, always ask yourself some questions:

- Are individuals and groups represented accurately?
- Are stereotypes relied upon?
- Are the outputs improperly biased in any way?

Don't assume that people are represented accurately; always consider your assumptions and check. With GenAI systems increasingly facing government regulation to ensure

accuracy and lack of bias, doing your own due diligence can help you to avoid relying upon stereotypes and falling afoul of legislation.

Knowledge Check: Ethical Risks of AI

1. When using GenAI, it's important to take ethics into account. Select the ethical issues to consider when using GenAI.
 - A. Are individuals and groups represented accurately?
 - B. Are stereotypes relied upon?
 - C. Are the outputs improperly biased?
 - D. Does the output seem plausible to me?
 - E. Does the content surprise me?

Answer: A, B, C

Explanation:

- A. This option is correct. AI can create inaccurate content about people or groups of people, so it's essential to check its output.
- B. This option is correct. It's important to check whether AI reproduces improper stereotypes in its output.
- C. This option is correct. AI is trained on real-world data, meaning it can rely on biases contained in the dataset.
- D. This option is incorrect. GenAI by its design sounds plausible, so you need to consider your own biases before simply accepting its plausibility.
- E. This option is incorrect. Factual information may surprise you, based on your previous knowledge, but whether or not the information seems surprising doesn't determine whether it's true or ethical.

2. A manager uses ChatGPT to create a job posting for a senior management role in engineering. Throughout the posting, the candidate is referred to using male pronouns. Select the ethical considerations that have been violated in this case.

- A. Are stereotypes relied upon?
- B. Are the outputs improperly biased?
- C. Are individuals or groups represented fairly?
- D. Has an individual been defamed?
- E. Is the output hallucinated?

Answer: A, B, C

Explanation:

- A. This option is correct. GenAI may be reinforcing the stereotype that men are more suited for engineering and leadership roles.
- B. This option is correct. GenAI may be reproducing bias against women in engineering and leadership roles.
- C. This option is correct. In this instance, GenAI is incorrectly describing this role as being occupied by a male candidate, which is a potentially inaccurate representation.
- D. This option is incorrect. GenAI hasn't described a specific individual, so no defamation has happened.
- E. This option is incorrect. GenAI does sometimes hallucinate false information, but aside from the pronouns, there is no indication of any made-up information in the ad.

COURSE COMPLETION

Congratulations! This completes the Understanding AI Risks and Mitigations course.