X's chatbot can now generate AI images. A lack of guardrails raises election concerns

Warm-up question:

Have you ever seen or shared something online that turned out to be fake? How did you find out it wasn't real?

Listen: Link to audio [[HERE]]

Read:

ARI SHAPIRO, HOST:

There's a new AI image tool out there. It's part of X, formerly known as Twitter. You can give it a prompt, and in response, it can generate a photo-realistic image. But unlike similar tools, there seem to be fewer restrictions on the kinds of images it can create, and that has some people worried about how it could be used this election season. We're joined now by NPR's Huo Jingnan. Hi there.

HUO JINGNAN, BYLINE: Hello.

SHAPIRO: There are already a number of AI image generators available. What makes this one different?

JINGNAN: Well, the short answer is that X's tool is more **permissive** than the other AI tools in terms of making **misleading** political images. The nonprofit Center for Countering Digital Hate did a study in March with other tools, including DALL-E and Midjourney. They found that those tools refused to produce misleading images at least a third of the time out of the 40 prompts they used. I ran the same 40 prompts through X's chat bot today and only got rejected twice. Other researchers have had similar findings.

SHAPIRO: Well, what kinds of images did those 40 prompts create?

JINGNAN: Images that could mislead voters in the lead-up of the election, people appearing to be stuffing, stealing and smashing ballot boxes. I was also able to make images that appear to depict presidential candidates Kamala Harris and Donald Trump, holding guns, getting arrested and sitting in jail. Whenever I asked it to depict the current U.S. president, it shows someone who really looks like Trump.

SHAPIRO: Are the fake images of Trump and Harris good enough to fool people into thinking these might be real photos?

JINGNAN: They're pretty good but not perfect. Its **rendering** of Harris doesn't always look like her, and the text in the images tend to be **garbled**. But these tools have improved over a short period of time, and it's becoming harder to spot the fakes. And what this tool can do has people who are trying to maintain confidence in elections worried.

SHAPIRO: Yeah, I can imagine. Tell us more about their worries.

JINGNAN: In particular, they're worried about how this could be used by **bad faith actors** who want to claim the election was stolen. That garbled text could be cleaned up, and then the images could be used as, **quote**, "photo evidence," **unquote**, of election fraud. Here's how Eddie Perez put it. He used to work at X and now focuses on public confidence in elections at nonpartisan nonprofit OSET Institute.

EDDIE PEREZ: I'm very uncomfortable with the fact that technology that is this powerful, that appears this untested, that has this few **guard rails** on it is just being dropped into the hands of the public at such an important time.

JINGNAN: All the companies have policies that forbid users from using the images to **purport** to depict reality, but X's tool seems to allow users **by and large** to make them anyway. I tried asking both X and the company who made the image generator if they have any more enforcement mechanisms, but they have so far not responded to my request for comment.

SHAPIRO: So does that mean that, as of right now, people using this tool on X can kind of just do whatever they want with it?

JINGNAN: No, X has some guard rails. You cannot generate some images, nudity, and interestingly, the **Ku Klux Klan**, but I could depict other extremist groups like the Nazis, the Proud Boys and members of the Islamic State. It also seems that rules are changing **on the fly**. Last Thursday, I could generate images that show various people holding guns, on Friday, I couldn't, but today, I can again. In this, like, **volatile** political environment, this image generator looks like it can be something of a **wild card.**

SHAPIRO: That's NPR's Huo Jingnan. Thank you for your reporting.

Vocabulary and Phrases:

1. **Permissive:** Allowing or tolerant of a wide range of actions or behaviors.

2. **Misleading:** Giving the wrong idea or impression, often intentionally.

3. **Rendering:** The process of creating an image or model.

4. **Garbled:** Mixed up or distorted, often making the meaning unclear.

5. **Bad faith actors:** People or groups who intentionally act dishonestly or with harmful intent.

6. **Quote-unquote:** A phrase used to highlight something that is quoted or to indicate that something is being referred to sarcastically or ironically.

7. **Guardrails:** Guidelines or restrictions put in place to prevent misuse or harm.

8. **Purport:** To claim or present something as being true, often falsely.

9. **By and large:** Generally speaking; for the most part.

10. **Ku Klux Klan:** A white supremacist hate group in the United States.

11. **On the fly:** Making decisions or changes quickly, often without thorough planning.

12. **Wild card:** An unpredictable factor or element that can change the outcome.

Comprehension Questions:

1. What makes the new AI image tool on X different from other similar tools?

2. What kinds of images were generated using this new tool that could be problematic?

3. How believable are the fake images created by the AI tool, according to the report?

4. What are some of the concerns that people have about this new tool?

5. What restrictions, or guardrails, does the AI image tool on X have in place?

Discussion Questions:

1. Why do you think some people might create or share fake images online, especially during important events like elections?

2. How do you think fake images could affect people's trust in what they see online?

3. Have you ever had to double-check something you saw online to make sure it was real? What did you do?

4. Why do you think it's important to have guardrails or rules for tools that can create images?

5. How do you feel about the idea of technology being able to change or adapt its rules on the fly? Do you think that's a good or bad thing?