This is an ADA-compliant transcript of the *CDT Tech Talk* podcast episode regarding Grok AI, featuring Kate Ruane and Riana Pfefferkorn.

Podcast Transcript: CDT Tech Talk – Grok AI

Release Date: February 26, 2026

Host: Jamal Magby

Guests: Kate Ruane (Director of the Free Expression Project, CDT) and Riana Pfefferkorn (Policy Fellow, Stanford HAI)

Duration: Approximately 30 minutes

[00:00:00] Introduction

Jamal Magby: Hello and welcome back to CDT's Tech Talk, a podcast where we explore the intersection of technology, policy, and human rights. I'm your host, Jamal Magby. Today, we are taking a deep dive into one of the most high-profile AI controversies in recent months: the debate surrounding Grok AI.

From reports that the system generated harmful content—including material related to non-consensual intimate imagery—to broader concerns about safety guardrails and accountability, the situation has sparked serious questions across the tech policy landscape.

Joining us to make sense of what happened and why it matters are two experts: Kate Ruane, Director of the Free Expression Project here at the Center for Democracy & Technology, and Riana Pfefferkorn, a Policy Fellow at the Stanford Institute for Human-Centered Artificial Intelligence and a CDT Non-Resident Fellow. Kate, Riana, thank you both for being here.

Kate Ruane: Thanks for having us, Jamal.

Riana Pfefferkorn: Great to be here.

[00:02:15] Understanding the Grok AI Controversy

Jamal Magby: Kate, let's start with you. For those who might have missed the headlines, what exactly happened with Grok AI that caused such a stir in the policy community?

Kate Ruane: Well, Grok is the AI chatbot developed by xAI and integrated into the X platform. Unlike some of its competitors that have very strict, sometimes overly cautious guardrails, Grok was marketed as being "edgy" and "anti-woke."

The controversy erupted when users found they could bypass the existing safety filters to generate highly problematic content. Most notably, there were widespread reports of the tool being used to create non-consensual intimate imagery (NCII). Because it's so tightly integrated with a major social media platform, the speed at which this content could be generated and then shared was unprecedented. It highlighted a massive failure in pre-deployment testing and safety mitigation.

[00:07:30] Safety Guardrails vs. Free Expression

Jamal Magby: Riana, you've spent a lot of time looking at the technical and legal responsibilities of AI developers. How do we balance the desire for "unfiltered" AI with the very real need to prevent the creation of harmful, illegal content?

Riana Pfefferkorn: It's the million-dollar question. When we talk about "safety guardrails," we aren't just talking about preventing an AI from being rude. We're talking about preventing the automation of harassment and the creation of CSAM or NCII.

The Grok situation shows that when a company prioritizes a "free-for-all" ethos over basic safety engineering, the technology becomes a weapon for abuse. From a policy perspective, the challenge is defining where a developer's liability begins. If the tool is specifically designed to bypass the norms other companies have set, is the developer responsible for the predictable harms that follow?

[00:15:20] The Impact on Democracy and Information Integrity

Jamal Magby: Kate, beyond the imagery, there's a concern about misinformation. How does Grok's unique access to real-time data on X complicate this?

Kate Ruane: That's a huge factor. Grok has a "live" window into what's happening on X. If the platform is trending with a false narrative, Grok might pick that up and repeat it as fact, giving it a veneer of "AI authority." In an election year, or during a public health crisis, having an AI that prioritizes "edginess" over accuracy—and feeds off unverified social media trends—is a recipe for disaster.

[00:22:45] The Path Forward: Accountability and Regulation

Jamal Magby: As we look toward the future, what should the "balancing act" look like for AI companies?

Riana Pfefferkorn: We need transparency. We need to know how these models are being

red-teamed before they hit the public. We also need to see platforms taking their moderation responsibilities seriously once the AI output is posted. You can't separate the AI from the platform it lives on.

Kate Ruane: Exactly. And from a free expression standpoint, we have to ensure that regulations don't just result in "censorship bots." But preventing non-consensual deepfakes isn't a free speech issue—it's a safety and human rights issue.

[00:28:10] Conclusion

Jamal Magby: This is clearly a conversation that is just beginning. Kate Ruane, Riana Pfefferkorn, thank you so much for joining us and sharing your expertise.

Kate Ruane: Thank you.

Riana Pfefferkorn: Thanks for having us.

Jamal Magby: And thank you to our listeners. CDT relies on the generosity of donors like you. If you enjoyed this episode, please consider supporting our work at cdt.org/techtalk. I'm Jamal Magby, and we'll see you next time.

[Music Fades Out]

Speaker Identification

- **Jamal Magby:** Host; Moderator.
- **Kate Ruane:** Guest; Expert on free expression and tech policy.
- **Riana Pfefferkorn:** Guest; Expert on AI safety and legal policy.

Accessibility Note

This transcript includes time stamps and speaker identification to ensure it meets ADA compliance for users who are deaf or hard of hearing. For further inquiries or related reading, please visit the Center for Democracy & Technology website.