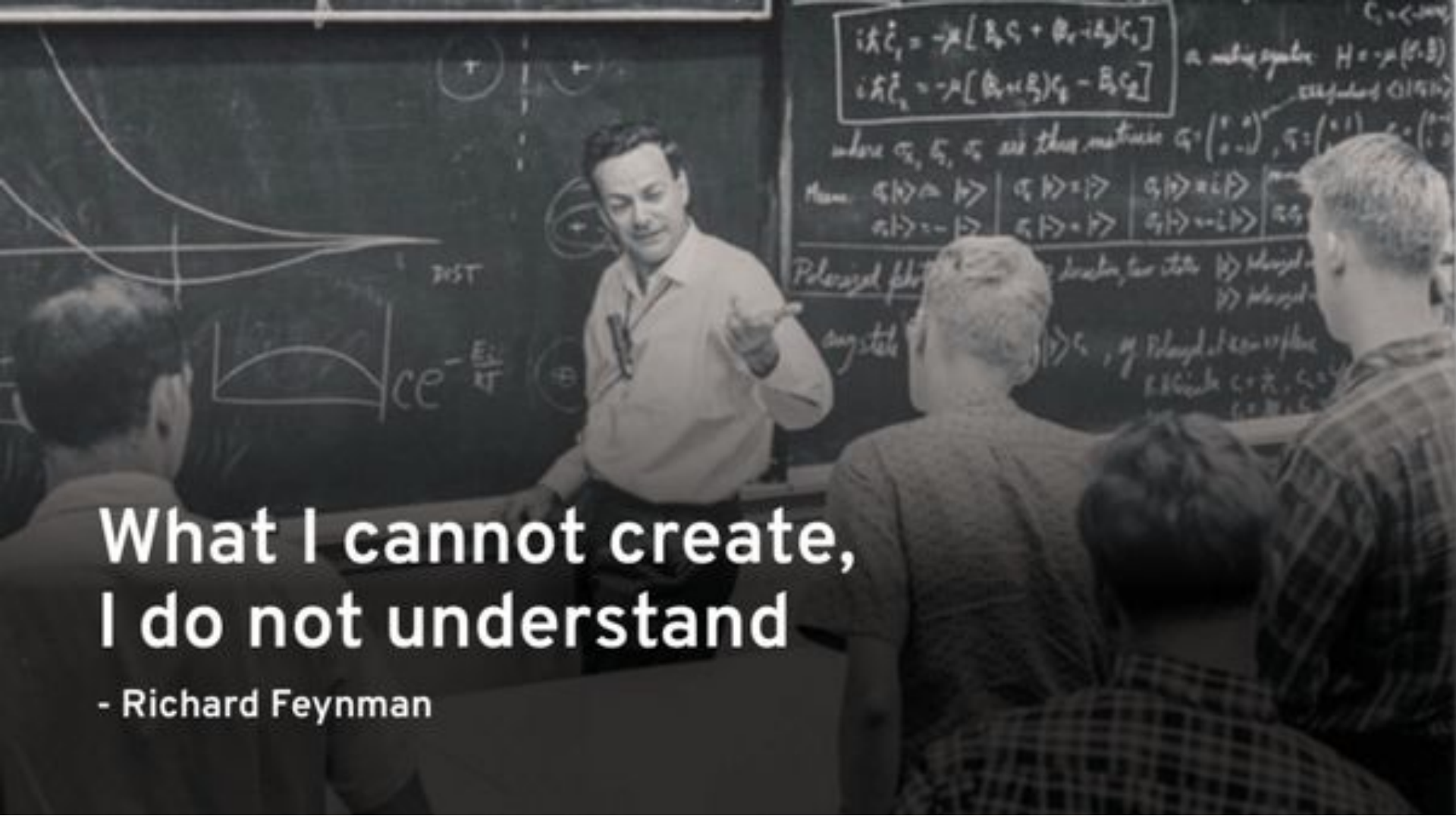
Countering AI Generated Disinformation on Social Media

Matthew Groh, Assistant Professor
Department of Management and Organizations, Kellogg School of Management
Department of Computer Science (by courtesy), McCormick School of Engineering
Core Faculty, Northwestern Institute on Complex Systems (NICO)

These images are AI generated



What I cannot create,
I do not understand

- Richard Feynman

Teach People the **Capabilities** and **Limitations** of Generative AI

By understanding the capabilities and limitations, people can (1) critically reason about the plausibility of an image, video, or audio clip based on the difficulty of producing such media (2) critically examine media for artifacts and plausibilities that continue to emerge even as media appears photorealistic on first appearance

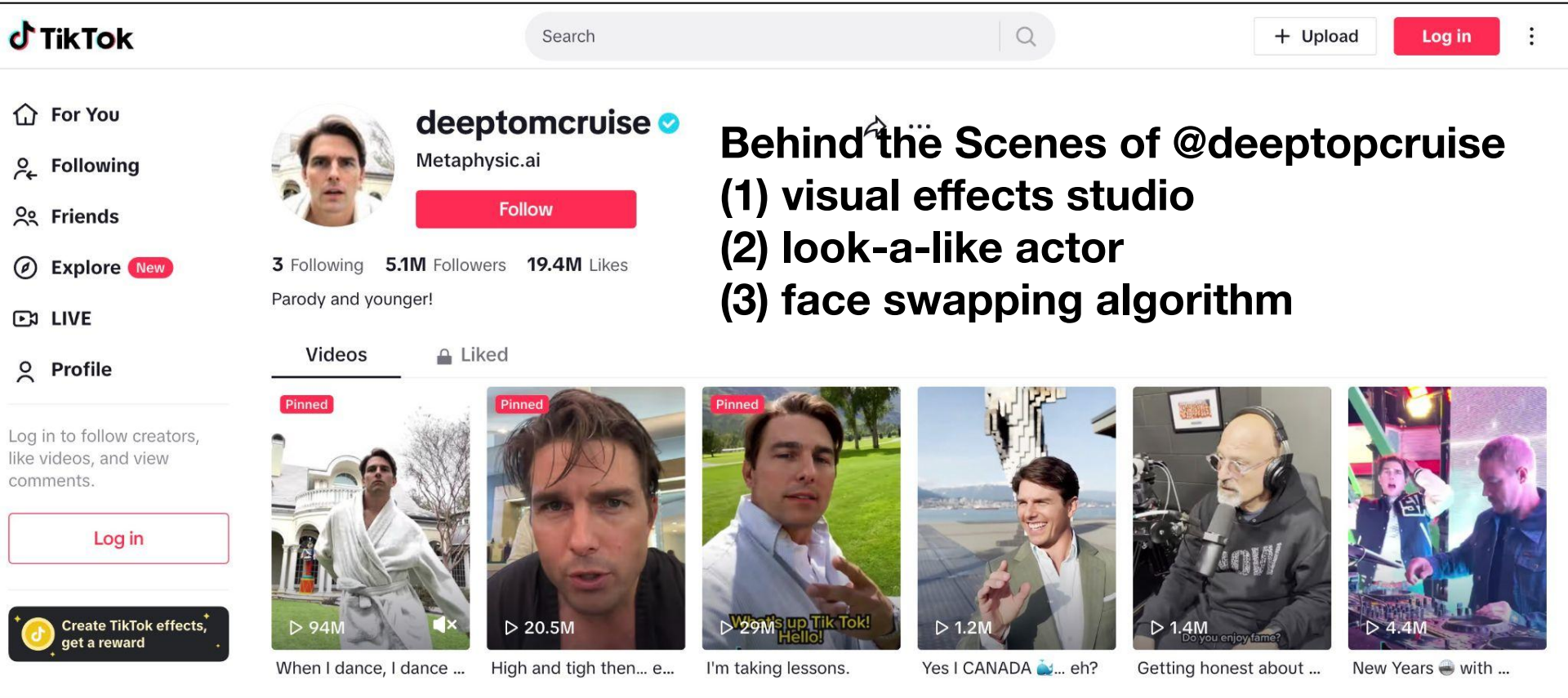
Generative AI literacy could extend the power of psychological inoculation, prebunking, and media literacy to today's changing technology landscape and will be most effective when paired with what makes generative AI persuasive, provocative, and likely to mislead.

Perceptual Realism Audits of Generative AI

Generative AI can produce photorealistic images, but to date, it is only possible in certain contexts (e.g. portrait photos that might appear in Linkedin but not group photos of friends hanging out eating pizza that you might see on Instagram). Moreover, these models only produce photorealistic images some of the time. Generative AI literacy will serve as an important complement to media literacy and psychological inoculation for countering disinformation.

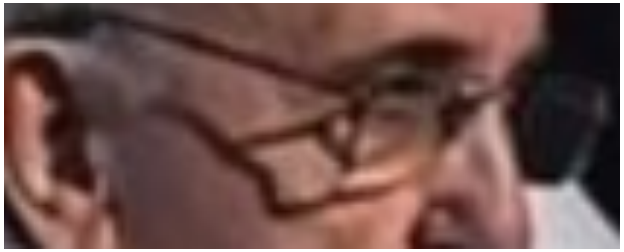
These images are AI generated

Tom Cruise deepfake is not just created by AI but instead by human artists making creative use of AI.

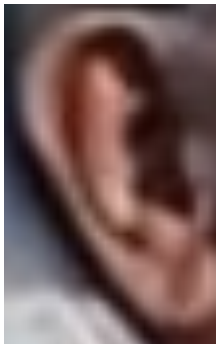


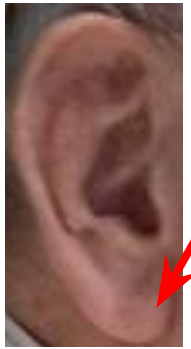
“Balenciaga Pope” looks real at first glance, but critical viewing reveals artifacts and implausibilities

Physical artifacts (shadow of glasses does not match rigidity of glasses)



Biometric artifacts (ears do not match)





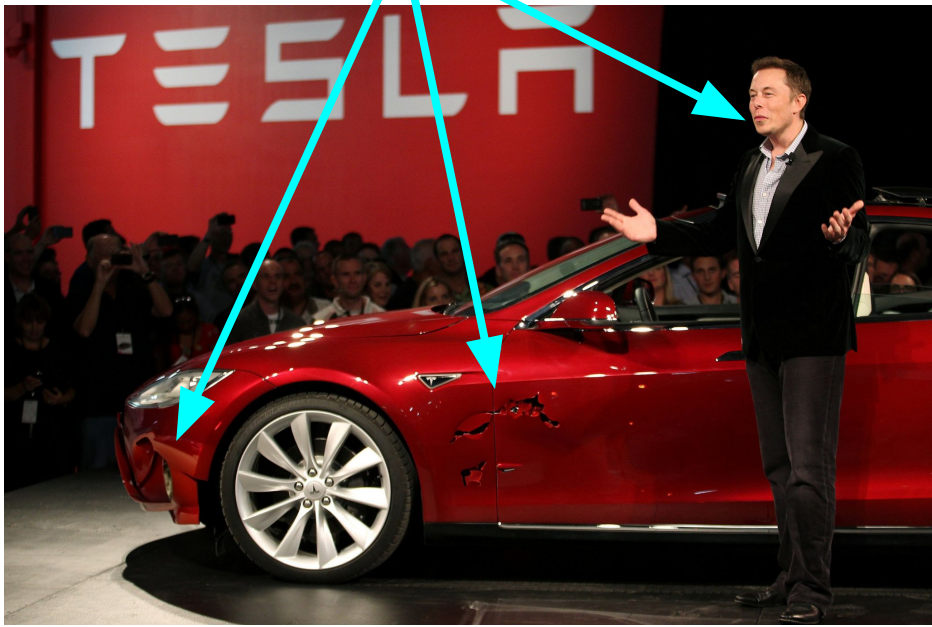
Verified photo of Pope

Functional implausibilities (the hand clasping the water bottle defies expectations of human grip and water bottle design)

We need to be careful about directing people's attention to the right artifacts.

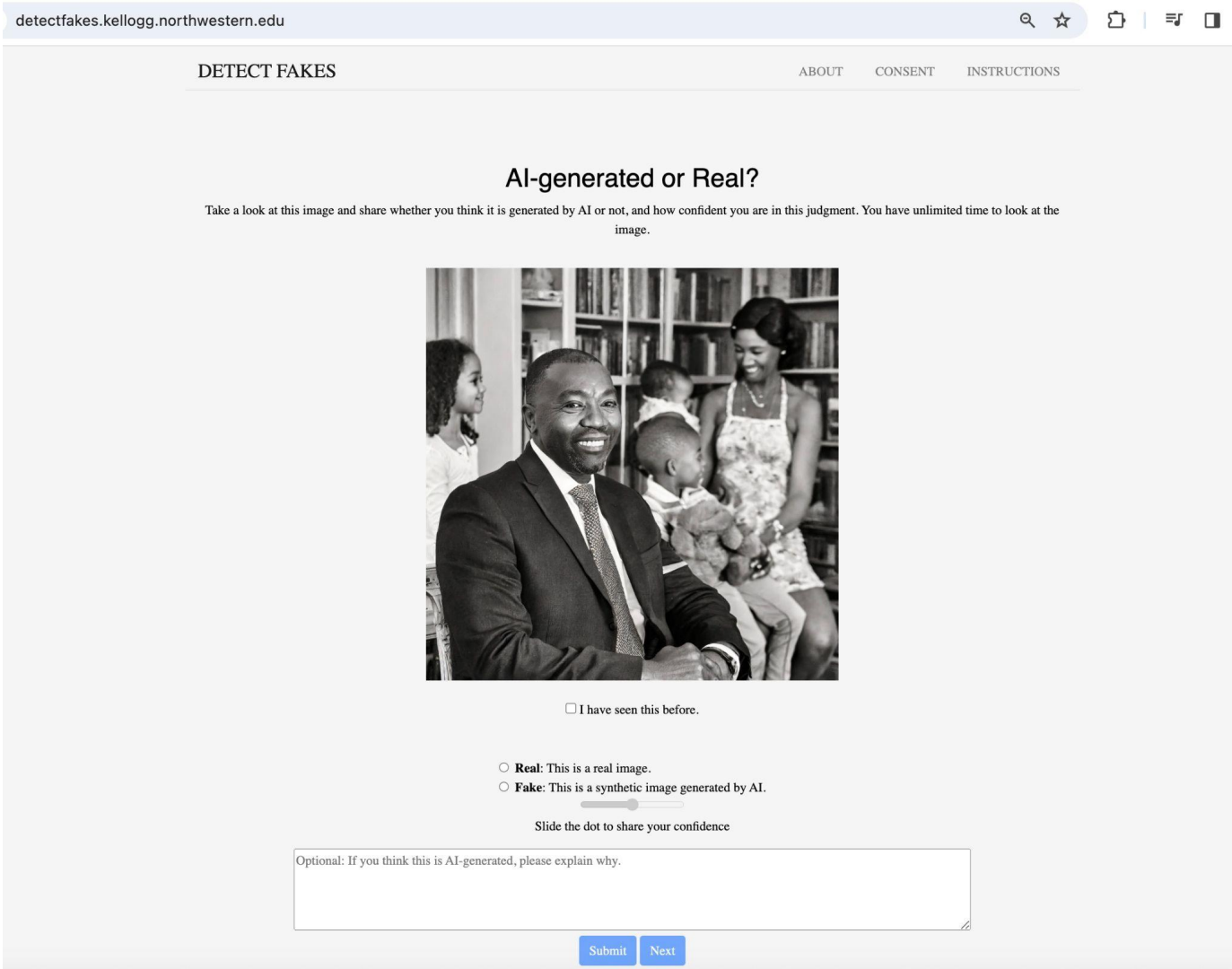
Otherwise, we can exacerbate the misleading effects of misinformation.

Some manipulations are simple and quick to create like these made with Adobe Firefly



Other manipulations are much harder to come by like these made with Midjourney with implausibilities around how pizza is held, drinks are placed, and fingers appear

Ongoing research evaluating when, where, why, and how ordinary people can spot AI-generated images and how to boost their abilities



Revelant Research on Deepfake Detection

- Human Detection of Political Deepfakes across Transcripts, Audio, and Video (2024) Matt Groh, Aruna Sankaranarayana, Nikhil Singh, Dong Young Kim, Andy Lippman, Rosalind Picard
- Art and the Science of Generative AI (2023) Ziv Epstein, Aaron Hertzmann, Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R Frank, Matt Groh, Laura Herman, Neil Leach, Robert Mahari, Alex Pentland, Olga Russakovsky, Hope Schroeder, Amy Smith
- Deepfake Detection by Human Crowds, Machines, and Machine-informed Crowds (2022) Matt Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard
- Human Detection of Machine-Manipulated Media (2021) Matt Groh, Ziv Epstein, Nick Obradovich, Manuel Cebrian, Iyad Rahwan

Example Outreach with Science Journal for Kids and Teens

How can humans and AI work together to detect deepfakes?

“What counts is not what sounds plausible, not what we would like to believe, not what one or two witnesses claim, but only what is supported by hard evidence rigorously and skeptically examined. Extraordinary claims require extraordinary evidence.”

— Carl Sagan