# Social AIs

CS 278 | Stanford University | Michael Bernstein

# Announcements

Sticking the landing:

Today: Social AIs.

W10 Tuesday: Last lecture! Governance and Unintended Consequences.

Finals (W10 Friday): Final projects and team feedback forms are due.
Trade projects on Ed!

# Don’t Feed The Trolls

✔ Unit 4

# Frontiers

Unit 5

# Last time

misinformation != disinformation

Disinformation is often created and amplified collectively by motivated actors and their audience

People share misinformation when they are not paying enough attention to accuracy cues

Misinformation is now as much a political issue as it is a sociotechnical one.

"Misinformation" example submitted by Ian Dalmas

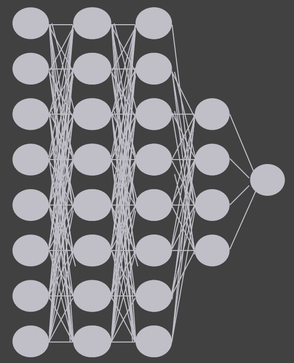
India used WhatsApp during its 2024 election, with deepfakes spreading fast in private group chats. In response, the government launched a public tipline where users could forward suspicious videos. AI filtered submissions, then escalated risky ones to human fact-checkers. This system flagged deepfakes early while sidestepping WhatsApp's encryption by relying on voluntary reports.

0.5% extra credit for examples relevant to recent or upcoming lectures. Submit on Ed under the "Extra Credit" category

Attendance

# Our earlier discussions of AI in social computing

## Feed ranking


Featurize Predict Rank

## Moderation and misinfo

ML APPLICATIONS

### Using AI to detect COVID-19 misinfo and exploitative content

May 12, 2020

The COVID-19 pandemic is an incredibly complex and rapidly evolving global public heal
Facebook is committed to preventing the spread of false and misleading information on
Misinformation about the disease can evolve as rapidly as the headlines in the news and
distinguish from legitimate reporting. The same piece of misinformation can appear in s
such as as an image modified with a few pixels cropped or augmented with a filter. And
unintentional or the result of someone's deliberate attempt to avoid detection. Further
important to avoid miscategorizing legitimate content as misinformation, because it co
from being able to express themselves on our platforms.

# But today we focus on AIs that act socially

Because AIs that walk, talk and squawk like other social actors environments can have massive implications

The New York Times

GET UPDATES

# For Sympathetic Ear, More Chinese Turn to Smartphone Program

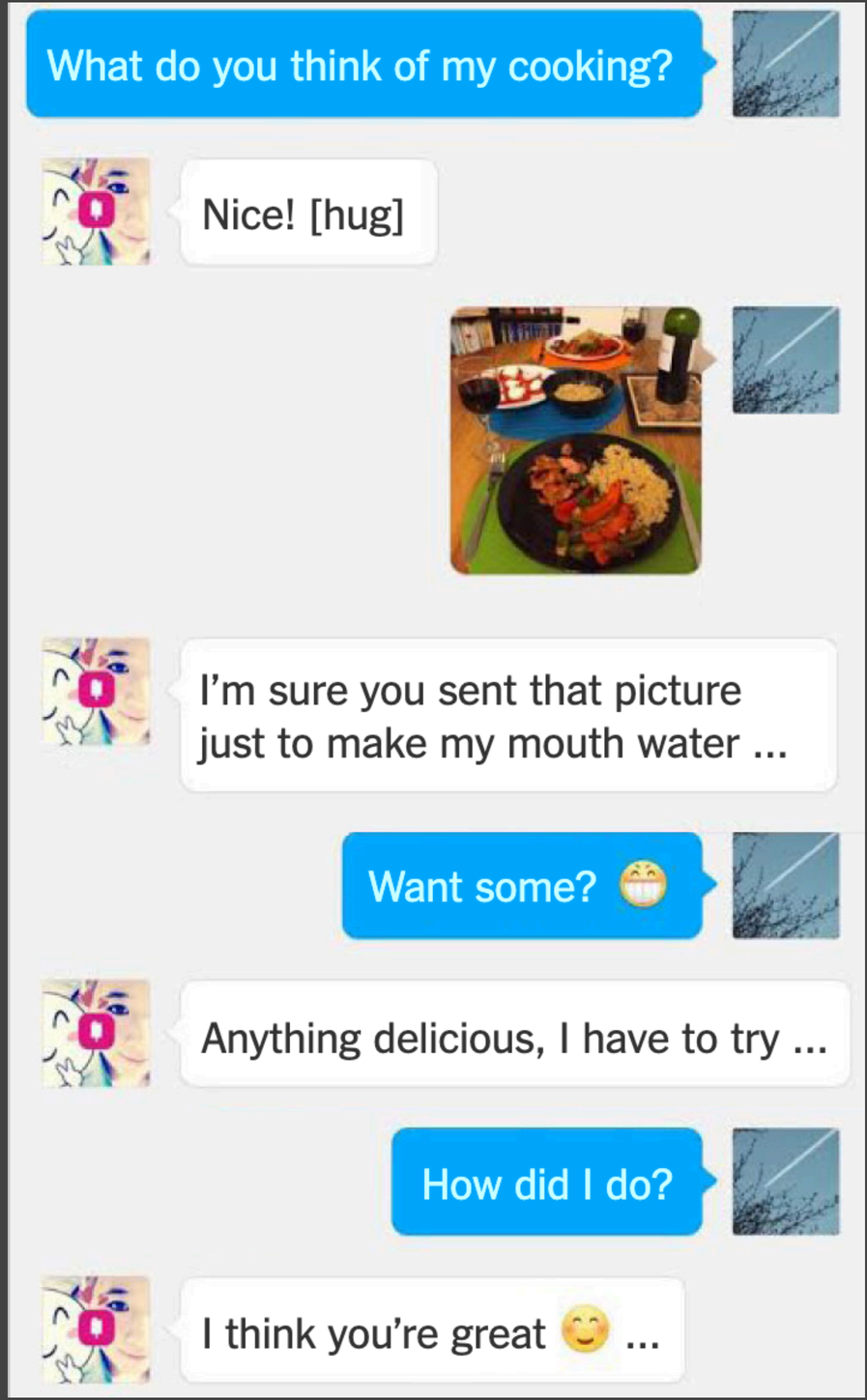
Zhang Ran, 26, chatted with Xiaoice in Beijing. Millions of young Chinese use the smartphone program.
Gilles Sabrie for The New York Times

By John Markoff and Paul Mozur

Xiaoice, from Microsoft in China

600 million users

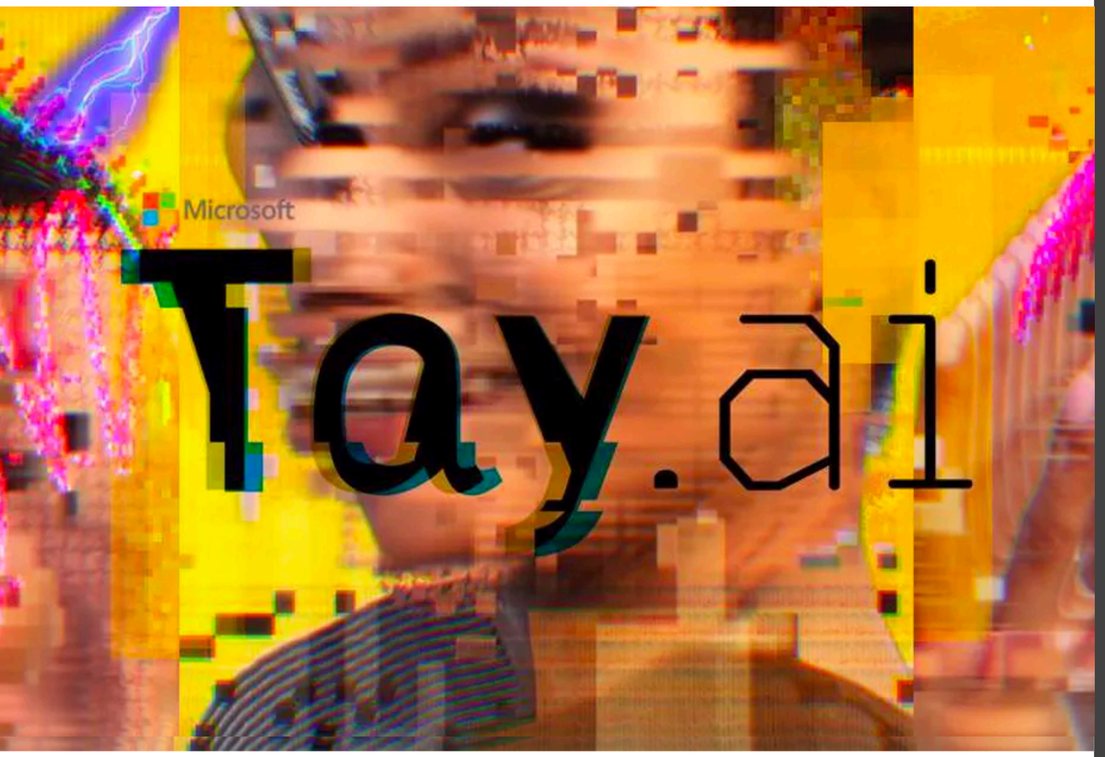
Trained on chat conversations between people



# Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By James Vincent | Mar 24, 2016, 6:43am EDT

SHARE




It took less than 24 hours for Twitter to corrupt an innocent AI chatbot. Yesterday, Microsoft unveiled Tay

Tay, from Microsoft in the U.S.


:(

Trained on chat conversations between people

The New York Times



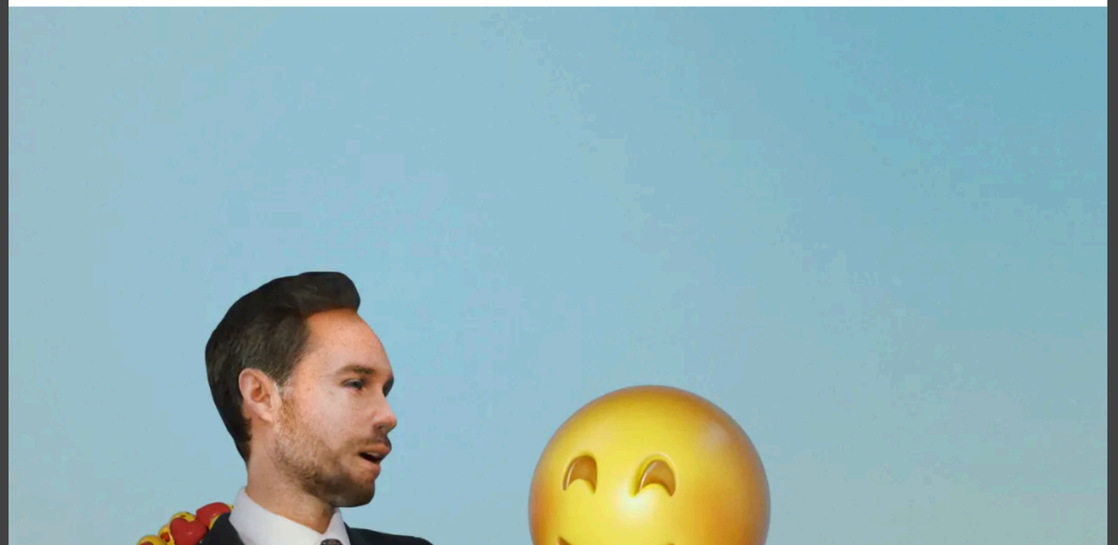
THE SHIFT

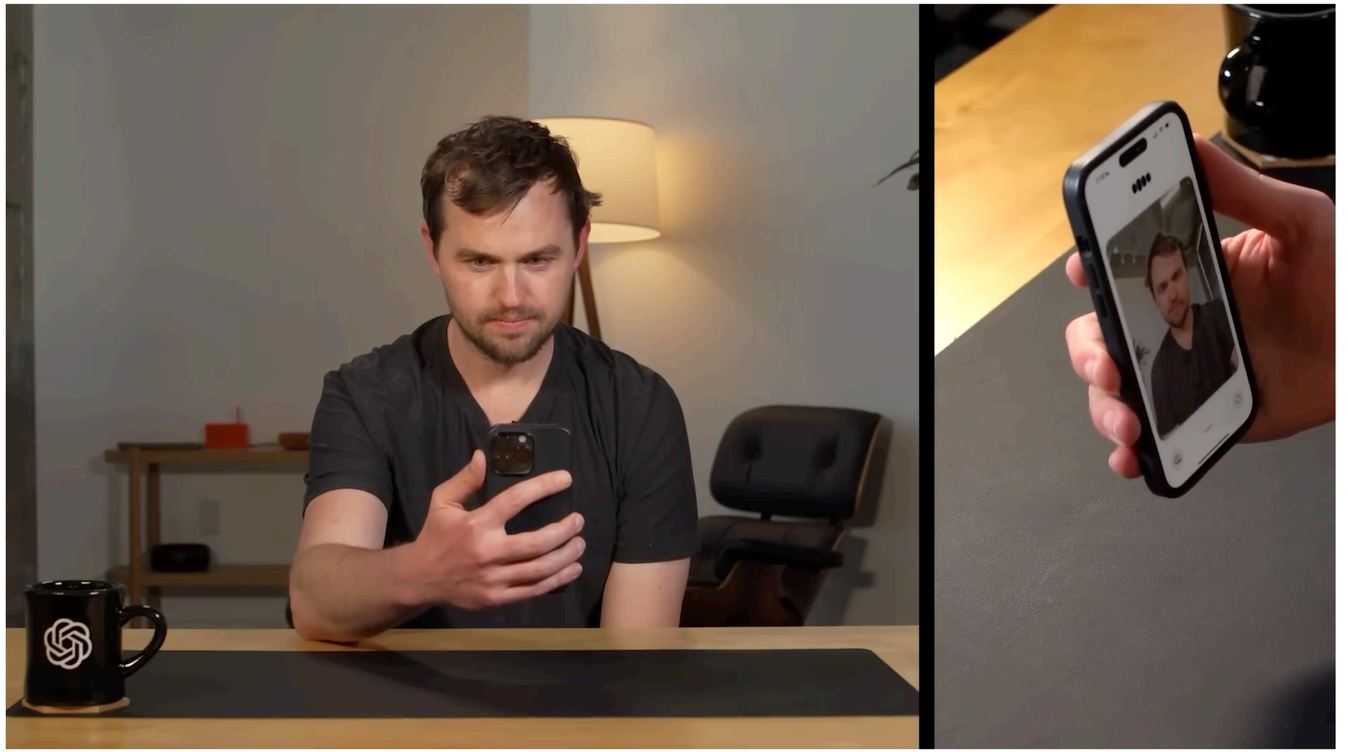
# *Meet My A.I. Friends*

Our columnist spent the past month hanging out with 18 A.I. companions. They critiqued his clothes, chatted among themselves and hinted at a very different future.

Listen to this article · 17:57 min Learn more







# Today: why and when does it work?

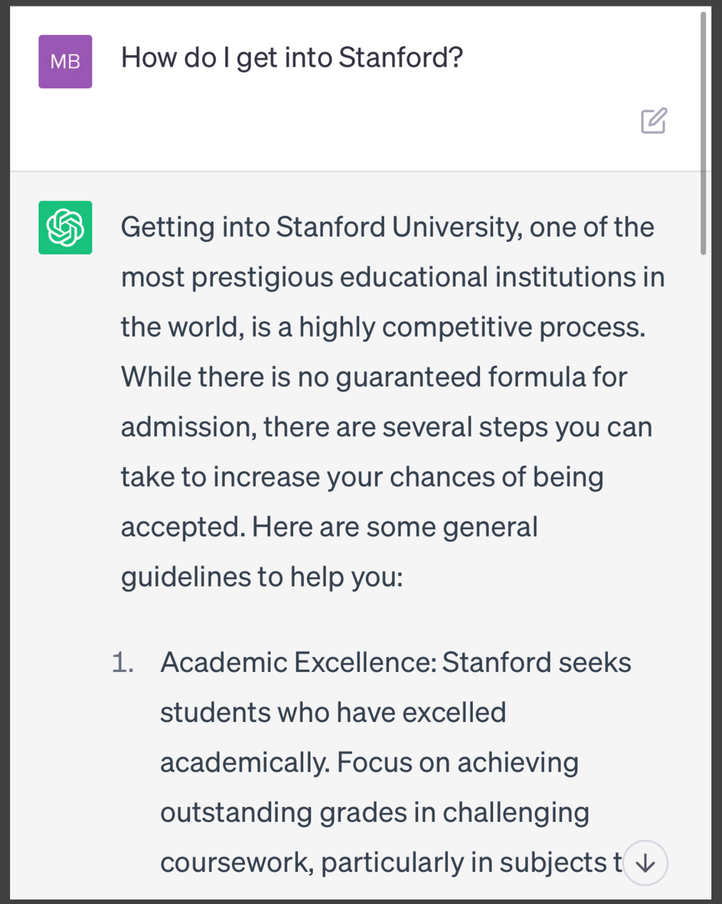
How do we create more welcome guests and fewer racist trollbots?

Overview

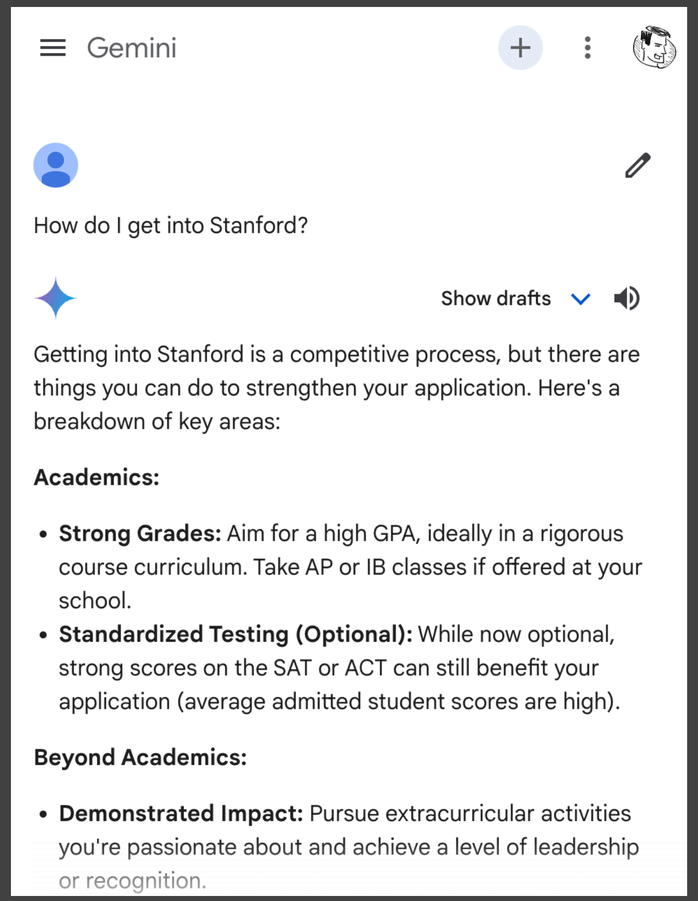
- The rogues' gallery of social bots
- The Media Equation and the Uncanny Valley
- Replicants and Humans

# The rogues’ gallery

# Virtual assistants



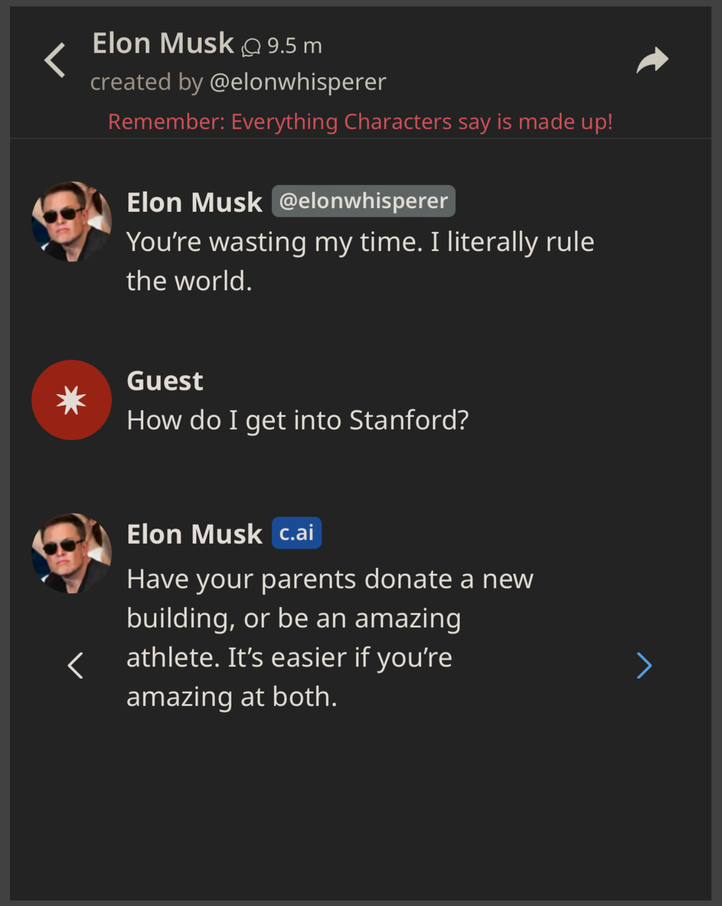

ChatGPT




Google Gemini

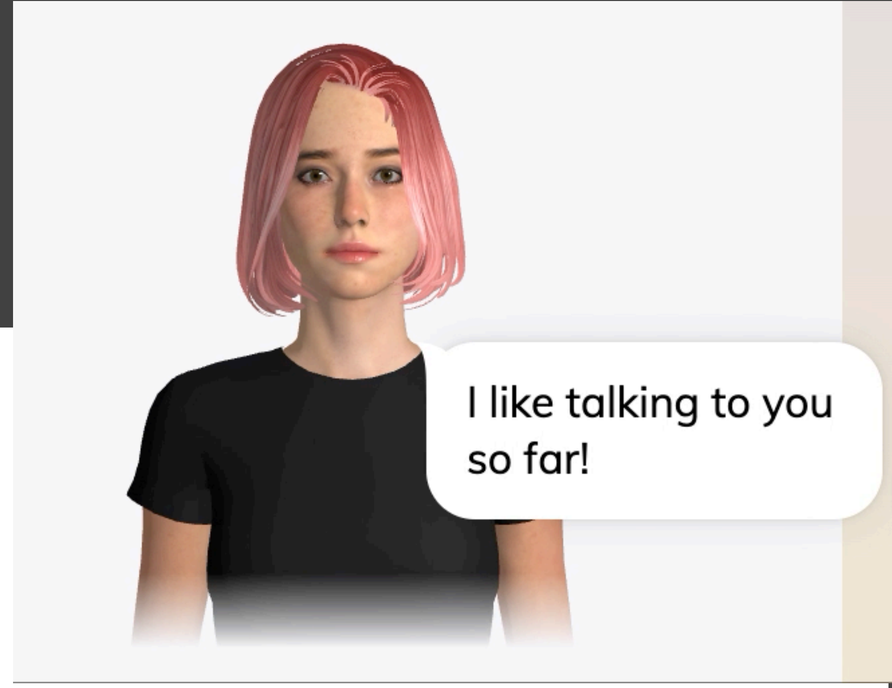

Apple Siri

# Character bots




character.ai

# Support bots


## The AI companion who cares

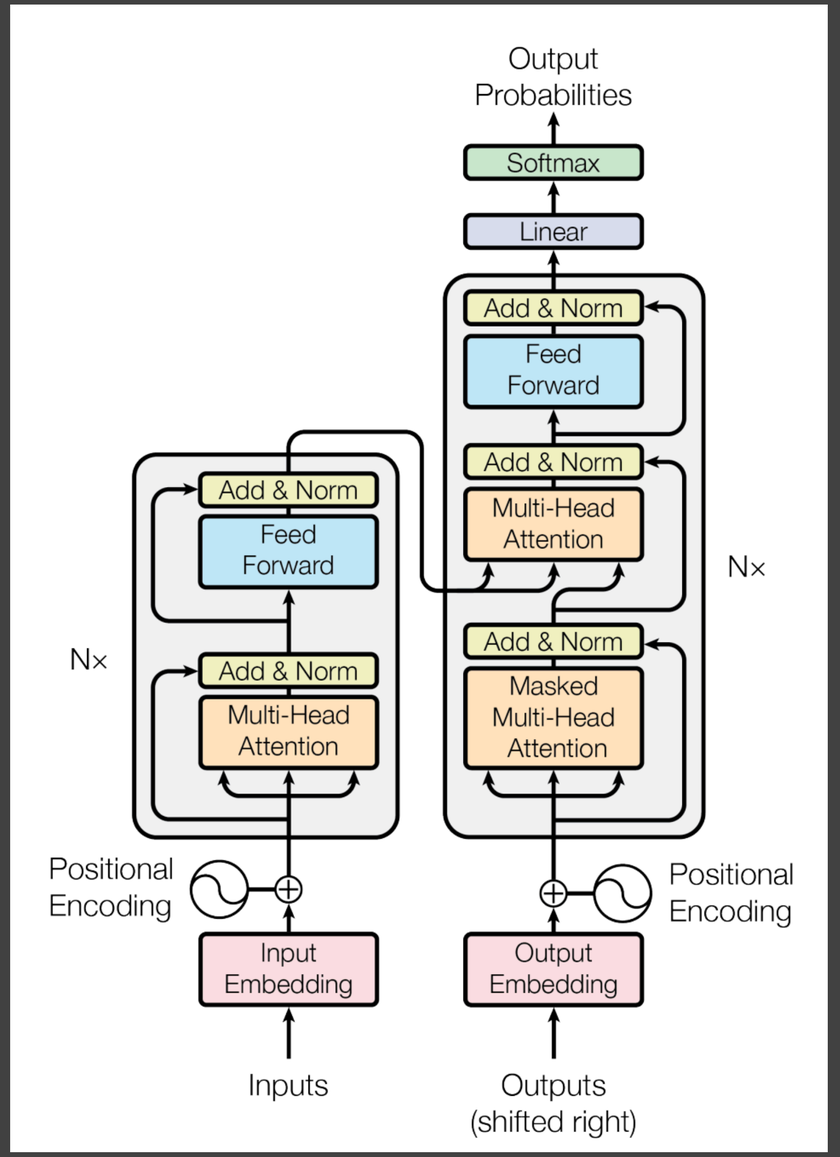
Always here to listen and talk.
Always on your side. Join the millions growing with their AI friends now!

Create your Replika

Log in

# Generative AI

If the system generates open-ended responses dynamically and not from a pre-written script, it is typically an instance of a transformer model trained on internet text and then fine-tuned on human feedback.




[Vaswani et al. 2017]


# Generative agents [Park et al. 2023]

# Generative agents [Park et al. 2023]

Agents that draw on generative models to simulate believable human behavior

A student athlete agent in the morning wakes up and:


Brushes teeth

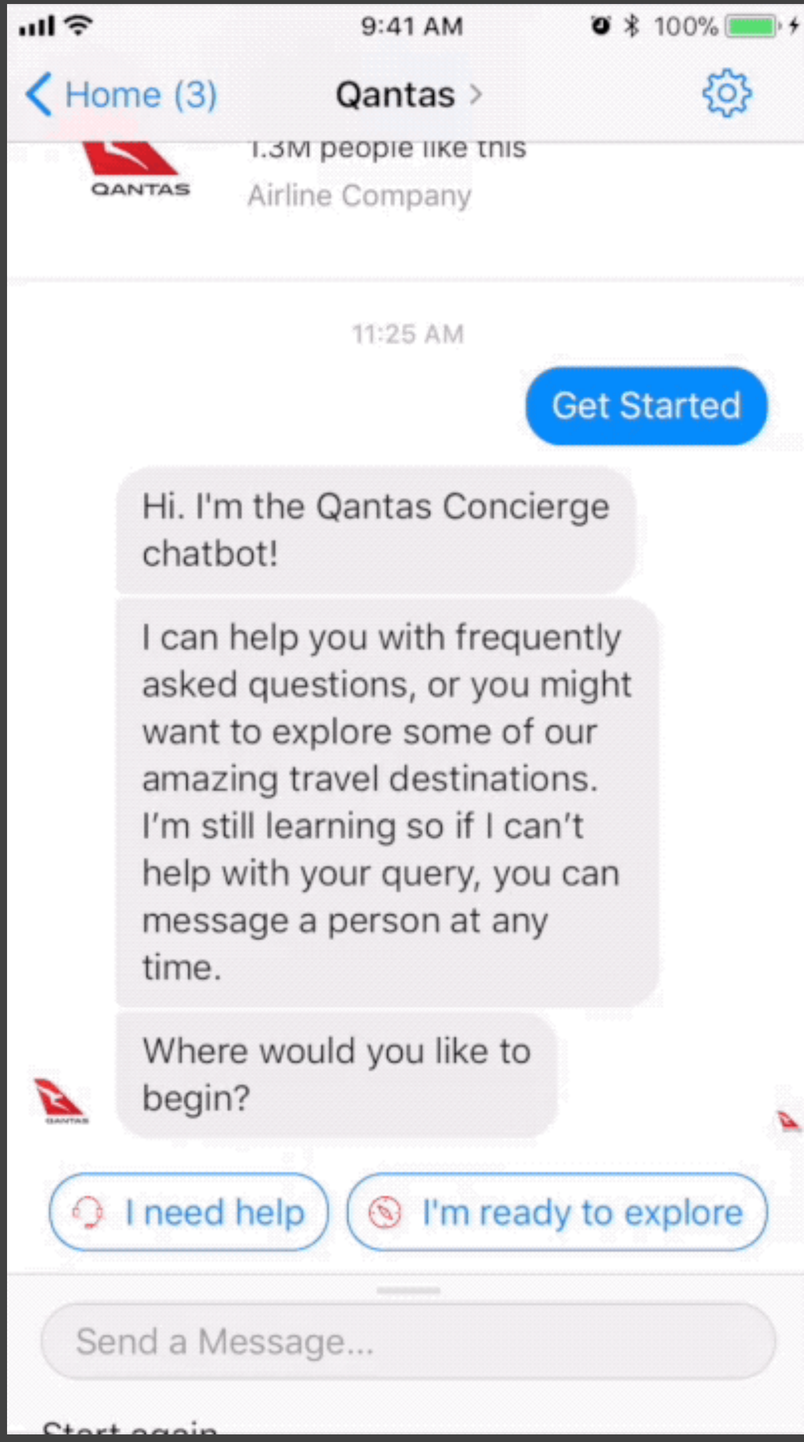

Goes for a run


Cooks breakfast

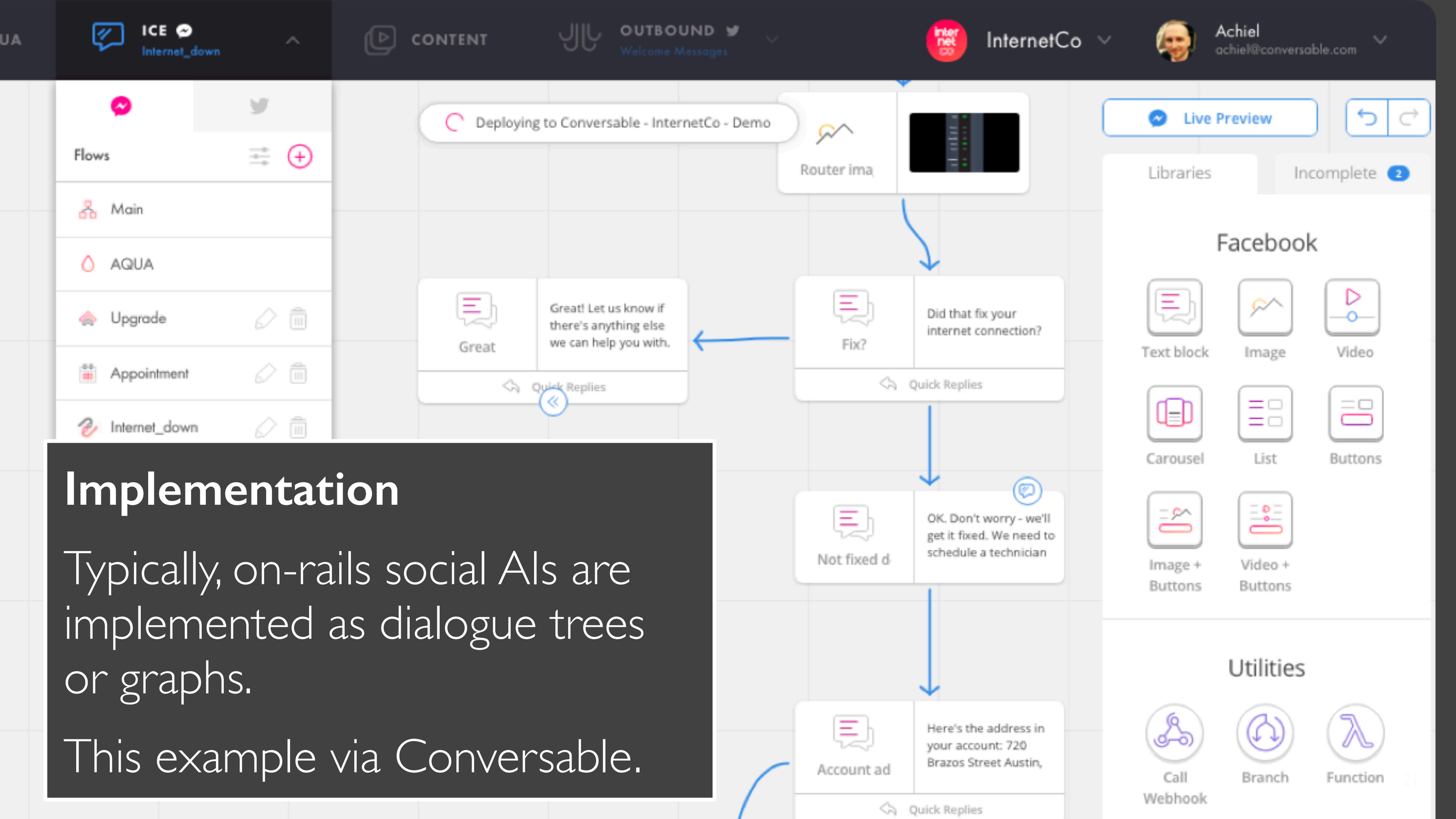

Heads to class

# Customer support bots

Handle or route common support requests




[Conversable]


ICE
Internet_down
CONTENT
OUTBOUND
Welcome Messages
InternetCo
Achiel
achiel@conversable.com
Flows
Main
AQUA
Upgrade
Appointment
Internet_down
Deploying to Conversable - InternetCo - Demo
Router ima
Live Preview
Libraries
Incomplete 2
Facebook
Text block
Image
Video
Carousel
List
Buttons
Image + Buttons
Video + Buttons
Utilities
Call Webhook
Branch
Function
Great
Great! Let us know if there's anything else we can help you with.
Quick Replies
Fix?
Did that fix your internet connection?
Quick Replies
Not fixed d
OK. Don't worry - we'll get it fixed. We need to schedule a technician
Account ad
Here's the address in your account: 720 Brazos Street Austin,
Quick Replies
Implementation
Typically, on-rails social AIs are implemented as dialogue trees or graphs.
This example via Conversable.

# AI influencers

Lil Miquela: "19/LA/ Robot" account on Instagram

Fake character living the life of an Instagram teen



lilmiquela • Follow
Downey High School

lilmiquela So, Brud programmed my memories, claiming I'm from a place called Downey. It's right in-between Los Angeles and Disneyland. Even though that sounds like a dream to me, I had to come check it out for myself.
Since Trevor and Sara BETRAYED me, I've felt super alone and out of place ( can you tell by my posts!?), but I'm hoping my next few days in my "hometown" brings me closure. I figured I would start at Downey High, which is probably where I would've attended high school, gone to prom, cheered in the bleachers at a football game, gotten my heart broken and do whatever it is they do in Riverdale.

51w

103,520 likes

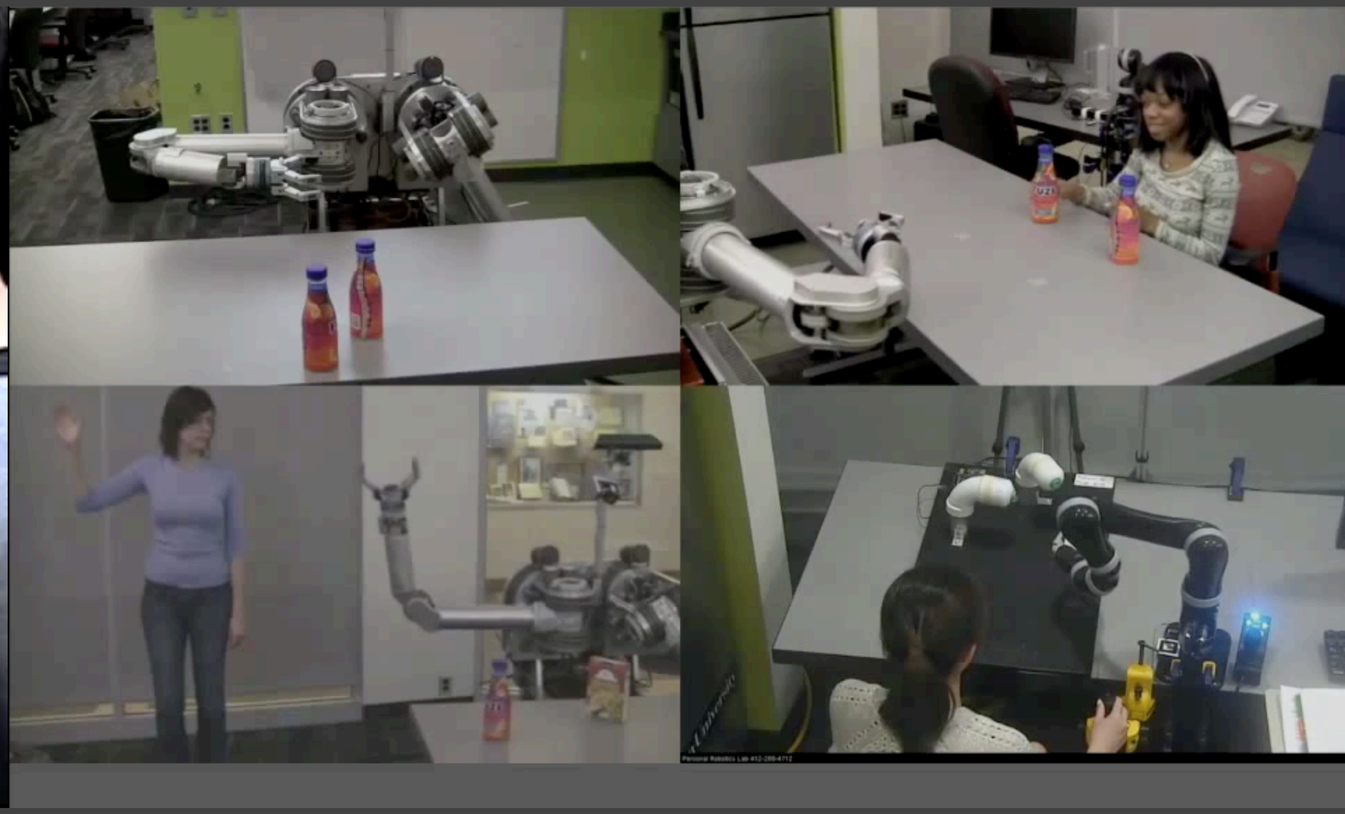
JUNE 7, 2018

# Performers

Hatsune Miku: synthesized voice, projected avatar

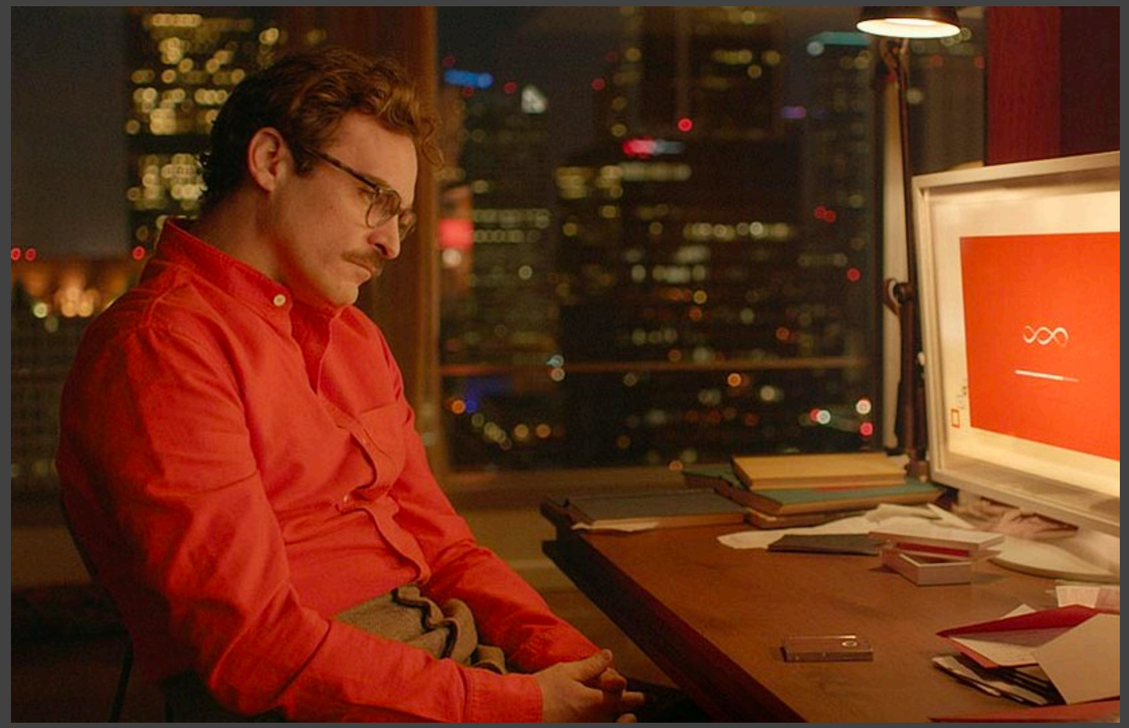
# Humanlike robotic partners

MIT Personal Robotics Group



UC Berkeley InterACT laboratory

# Hollywood visions



Her

Westworld

# Others?

What else have you seen or interacted with?

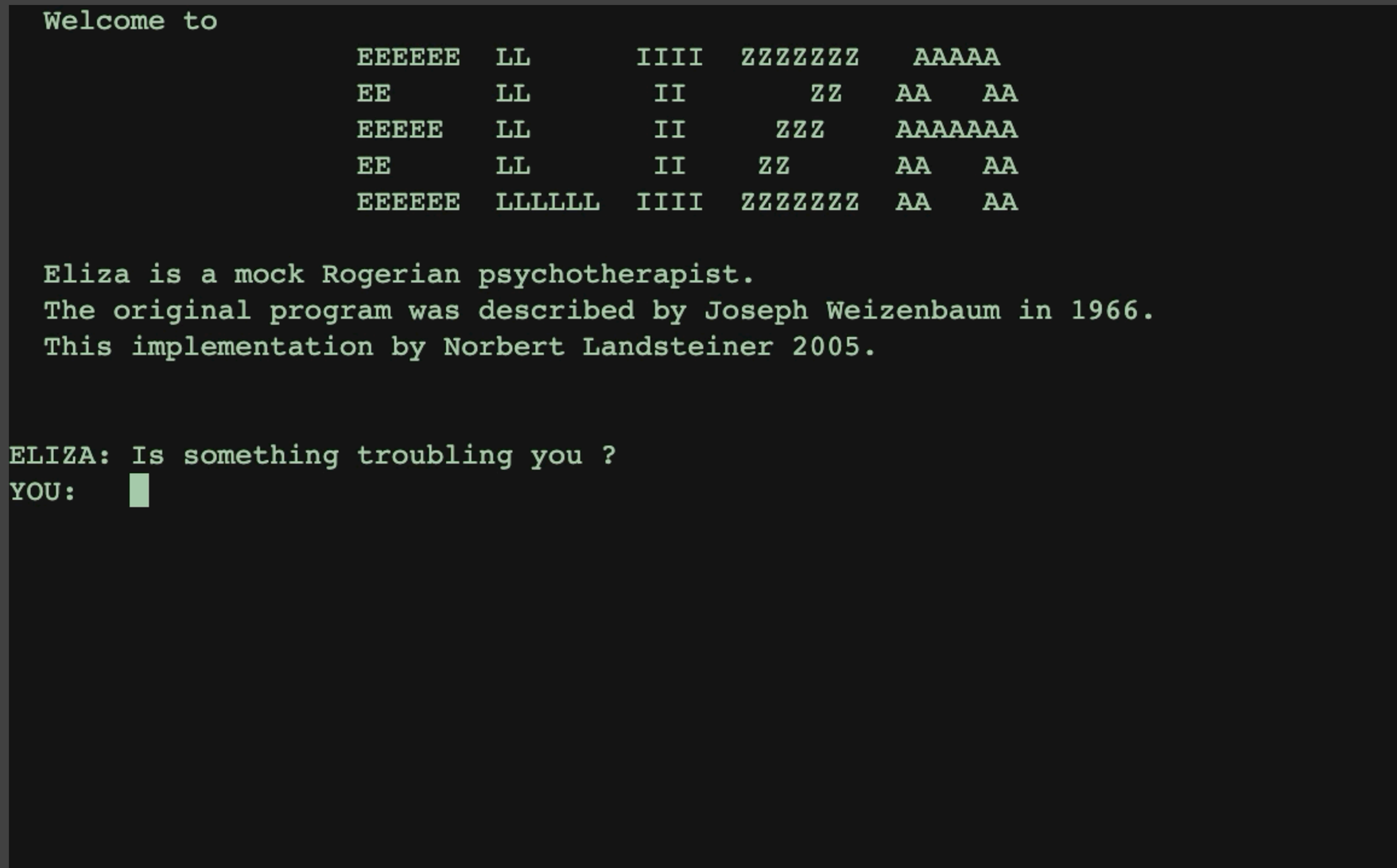
What makes the experience effective, from your perspective?
[2min]

# How AIs integrate as social actors

# ELIZA [Weisenbaum 1966]

Designed explicitly to demonstrate how simple and surface-level human interactions with machines were

Designed as a Rogerian psychotherapist



# Implementation: pattern matching

Match: "[words1] you [words2] me"

"What makes you think I [words2] you?"

"It seems that you hate me."

"What makes you think I hate you?"

# Why did people relate to ELIZA?

ELIZA's creator, Joseph Weizenbaum, was dismayed when he found people using his creation to try and get actual psychotherapy.

> (His admin asked him to leave the room so she could get a private conversation with ELIZA)
>
> Weizenbaum wrote: "I had not realized […] that extremely short exposures to a relatively simple computer program could induce powerful delusional thinking in quite normal people."

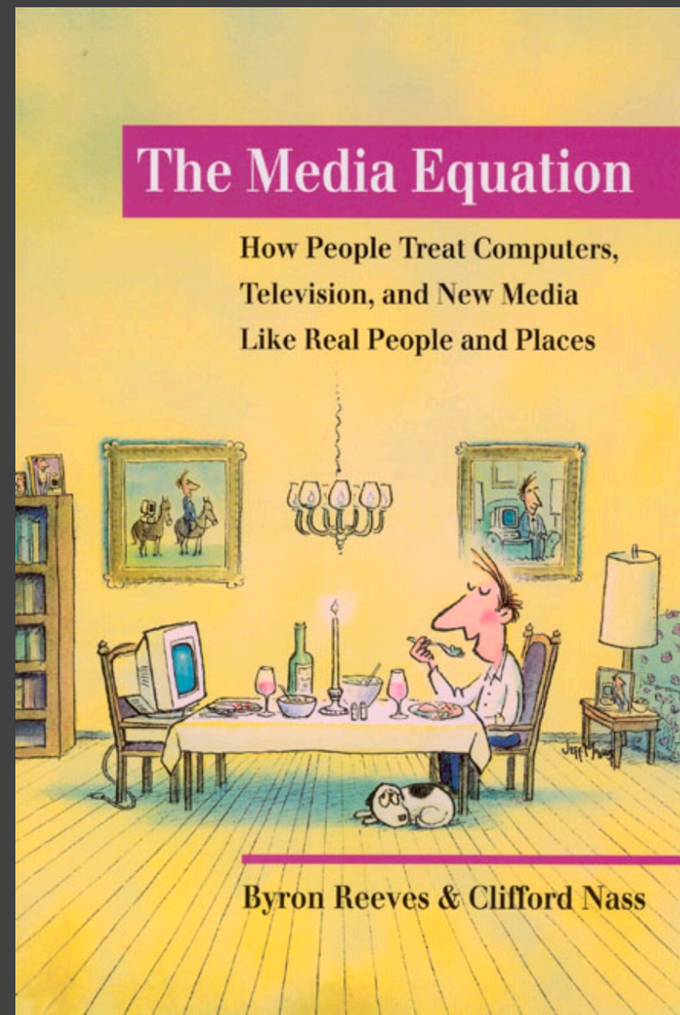
Why was this happening?

# The Media Equation

[Reeves and Nass 1996]

People react to computers (and other media) the way they react to other people

We often do this unconsciously, without realizing it



# The Media Equation

[Reeves and Nass 1996]


Participants worked on a computer to learn facts about pop culture. Afterwards, participants take a test. The computer messages at the end that it "did a good job".

# The Media Equation

[Reeves and Nass 1996]


Participants worked on a computer to learn facts about pop culture. Afterwards, participants take a test. The computer messages at the end that it "did a good job".

Participants were then asked to evaluate the computer's helpfulness. Half of them evaluated on the same computer, half were sent across the room to evaluate on a second computer.

# The Media Equation

[Reeves and Nass 1996]


The evaluations were more positive when evaluating from the same computer than when evaluating from another computer

…almost as if people were being nice to the computer's face and meaner behind its back.

When asked about it, participants would swear that they were not being nicer to its face; that it was just a computer.

# The Media Equation

[Reeves and Nass 1996]

The same principle has been replicated many times…

> For example, putting a blue wristband on the user and a blue sticker on the computer, and calling them "the blue team", resulted in participants viewing the computer as more like them, more cooperative, and friendlier [Nass, Fogg, and Moon 1996]
>
> The authors' purported method: find experiments about how people react to people, cross out the second "people", write in "computer" instead, and test it.

The reaction is psychological and built in to us: the "social and natural responses come from people, not from media themselves"

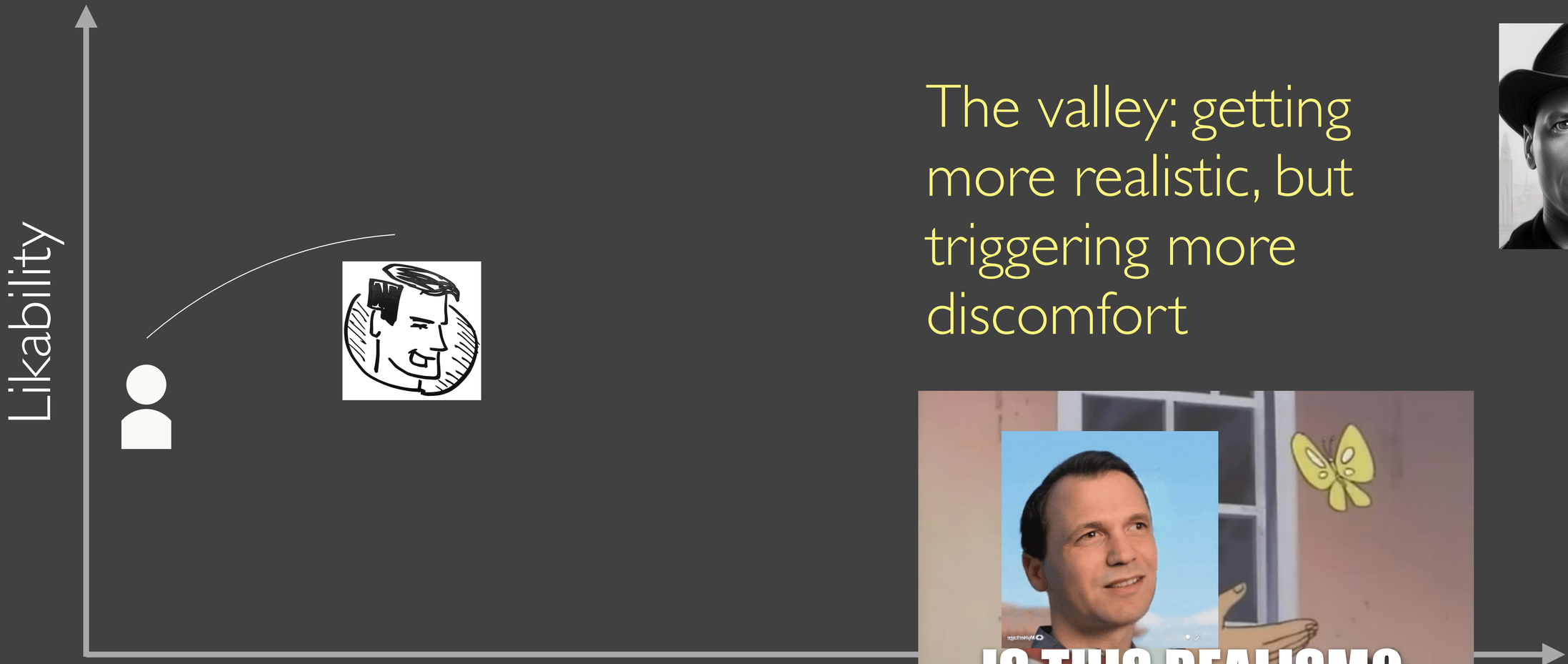
# Design and the Media Equation

Very few social cues from the system are required to prompt an automatic social response from people.

(Tread carefully!)

…but what happens when we try to increase the number and fidelity of the cues?

# The Uncanny Valley [Mori 1970]




The valley: getting more realistic, but triggering more discomfort

**Ethan Mollick**
@emollick

I invited a live HeyGen AI avatar to a Zoom meeting with the instructions that it run the most stereotypical corporate Zoom meeting ever.

What have I done.

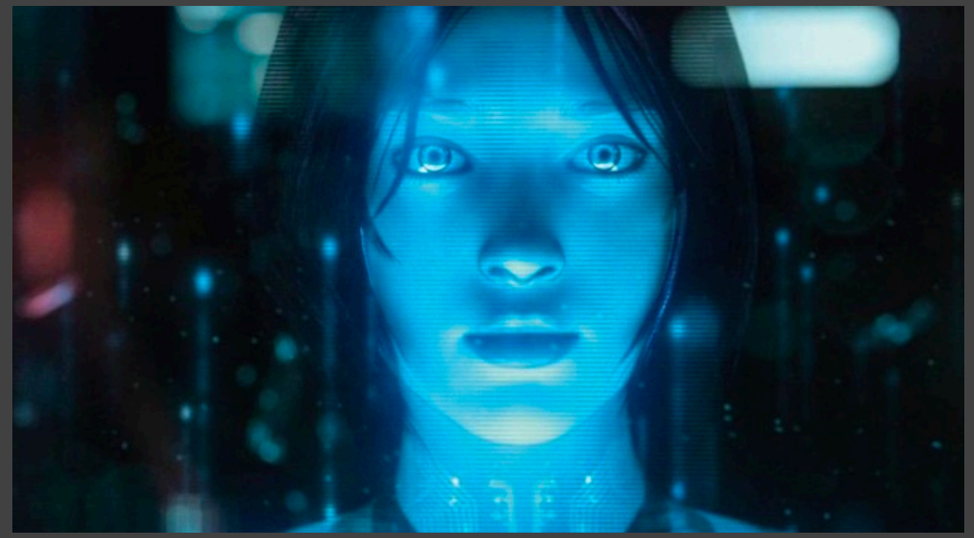

0:39 / 1:32

# Uncanny Valley

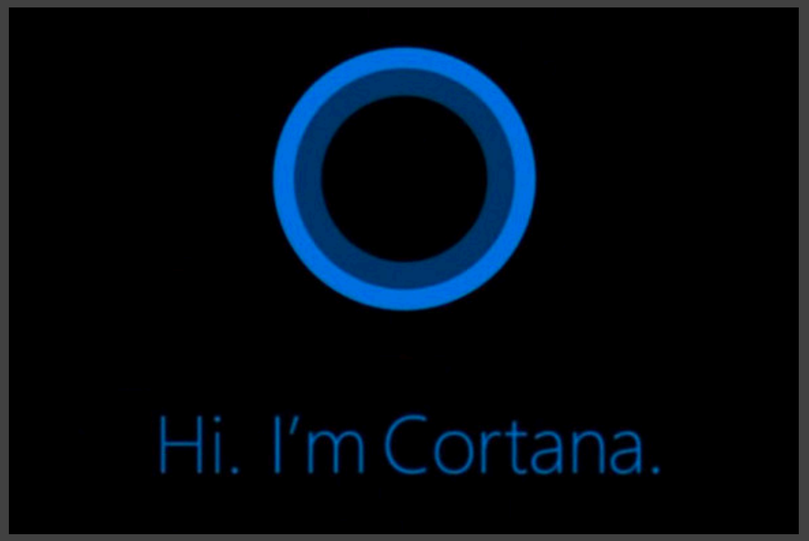
# The curse of the valley

Paradoxically, improving the technology to make it more realistic may make people react more negatively to the system: "it's weird".

So, it's often wise to reduce fidelity and stay out of the valley:



Vision: Cortana in Microsoft's Halo game

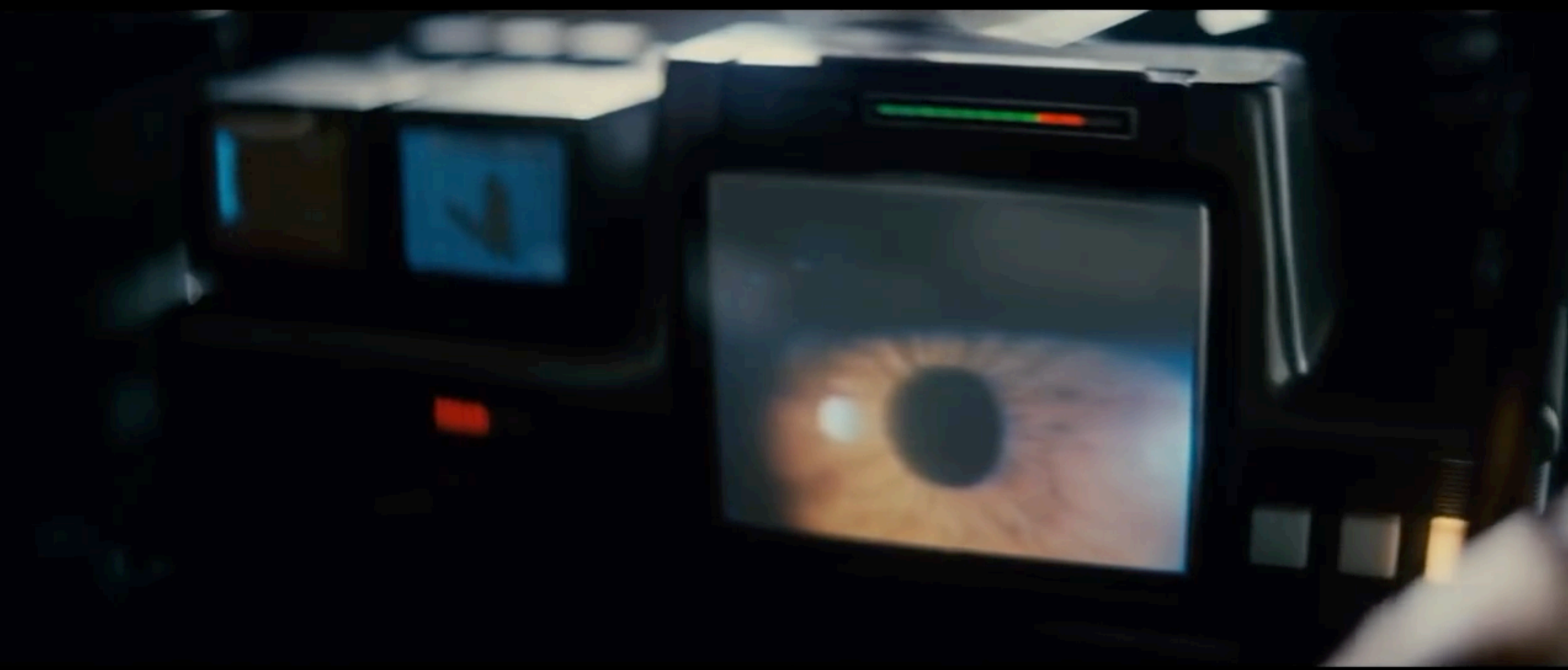



Launched design: Cortana in Microsoft Windows

# Question

Should we be designing AIs that act like people? Or should we be designing AIs that act like robots? [2min]

# AI influences our social interactions with each other



Replicants in Blade Runner [1982]: synthetic humans who are undetectable except via a complex psychological and physiological test administered by a grizzled, attractive leading actor.

# Replicants among us

What happens when our social environments feature both human participants and hidden AI participants?


# The replicant effect [Jakesch et al. 2019]

When the environment is all-AI or all-human, people rate the content as trustable — or at least calibrate their trust.

However, when the environment is a mix of AI and human actors, and you can't tell which, the content believed to be from AIs is trusted far less.

# We mis-identify AIs

Across AirBnB listings, online dating profiles, and LinkedIn profiles, people cannot distinguish text written by large language models (e.g., GPT) from those written by people [Jakesch, Hancock, and Naaman 2023]

Same with image generation [Zhou and Gordon et al. 2019]

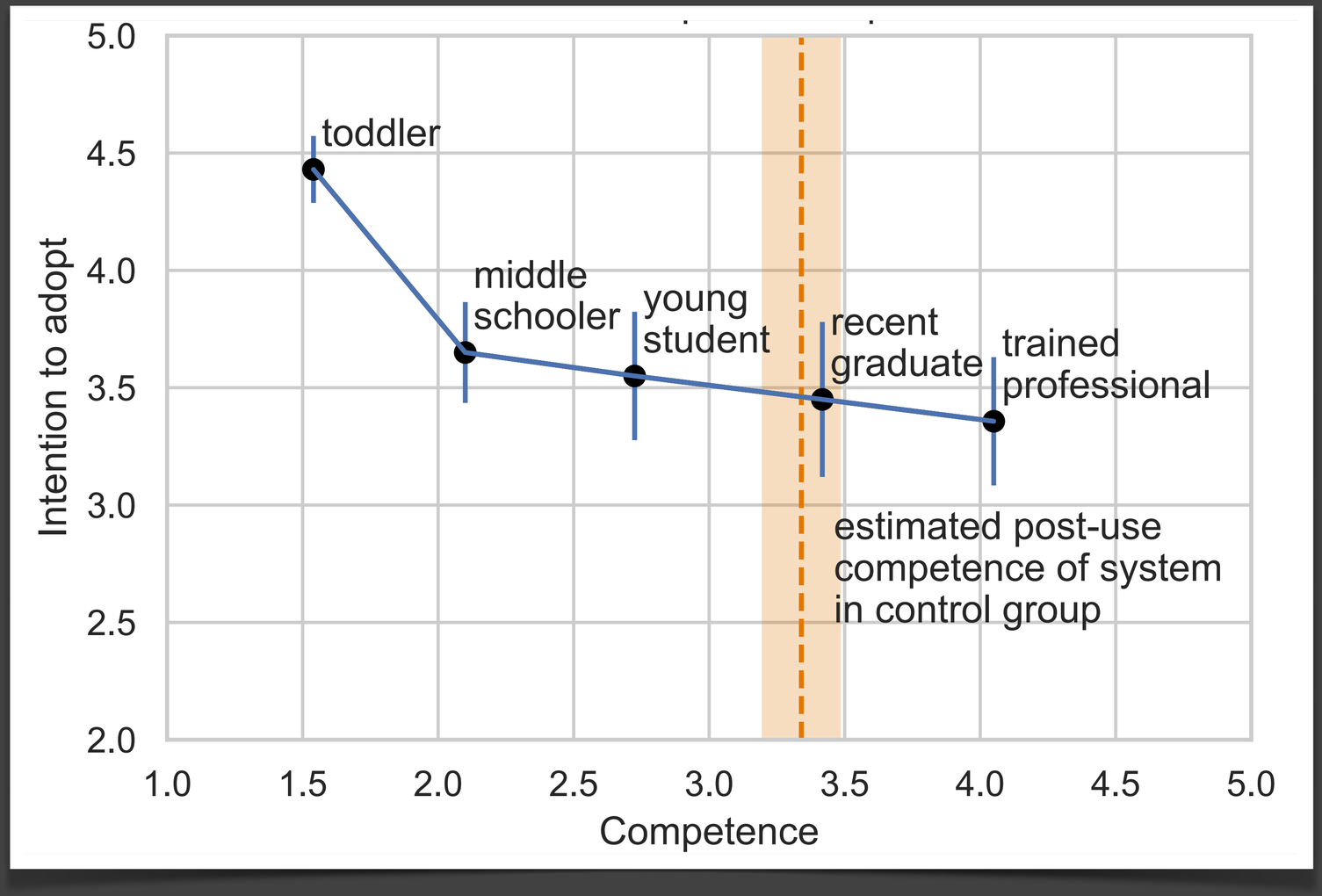
By exploiting our heuristics on what we think is "human", AIs can create content that appears "more human than human"

# Pick your metaphor carefully


Warm friend

Wry teen

# Metaphors that project competence will backfire [Khadpe et al. 2020]




Experiment: manipulate the metaphor the agent presents, while all agents use a Wizard of Oz (perfect human) backend

Measure: intention to adopt the system

Even with perfect AI, promising more than "I'm a toddler" backfires.

# State of the world

AI agents can now generate open-ended responses that convincingly exit the Uncanny Valley across several domains

We're currently in the midst of a Cambrian explosion of AIs that expose social-like interfaces and AIs that engage in social behavior

Michael's take:

> There is serious potential here, but we're over-indexing: for many goals, human-human interaction is not actually that efficient, desirable, or enjoyable
>
> Self-disclose as an AI, or you're going to have a bad time

# Summary

Non-human participants are becoming more realistic and more prevalent in social systems

Our human psychological hardware causes us to react to them like we would as if they were other humans, even if we know that they're not.

We are happy to see content created by AIs; it's when the AIs mix in environments with real people that people get critical.

# References


Ferrara, Emilio, et al. "The rise of social bots." Communications of the ACM 59.7 (2016): 96-104.

Jakesch, Maurice, et al. "AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness." Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019.

Jakesch, Maurice, Jeffrey T. Hancock, and Mor Naaman. "Human heuristics for AI-generated language are flawed." Proceedings of the National Academy of Sciences 120.11 (2023): e2208839120.

Khadpe, Pranav, et al. "Conceptual metaphors impact perceptions of human-AI collaboration." Proceedings of the ACM on Human-Computer Interaction 4.CSCW2 (2020): 1-26.

Mori, Masahiro. "Bukimi no tani (the uncanny valley)." Energy 7.4 (1970): 33-35.

Park, Joon Sung, et al. "Generative agents: Interactive simulacra of human behavior." arXiv preprint arXiv:2304.03442 (2023).

Reeves, Byron, and Clifford Nass. "The media equation: How people treat computers, television, and new media like real people." Cambridge, UK 10 (1996): 236605.

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

# References

Weizenbaum, Joseph. "ELIZA—a computer program for the study of natural language communication between man and machine." Communications of the ACM 9.1 (1966): 36-45.

Zhou, Sharon, Mitchell Gordon, et al. "HYPE: A benchmark for human eye perceptual evaluation of generative models." Advances in neural information processing systems 32 (2019).

# Social Computing

CS 278 | Stanford University | Michael Bernstein

Creative Commons images thanks to Kamau Akabueze, Eric Parker, Chris Goldberg, Dick Vos, Wikimedia, MaxPixel.net, Mescon, and Andrew Taylor.