



Embedded Ethics:

CSC401:
Anthropomorphization
(Module 2)

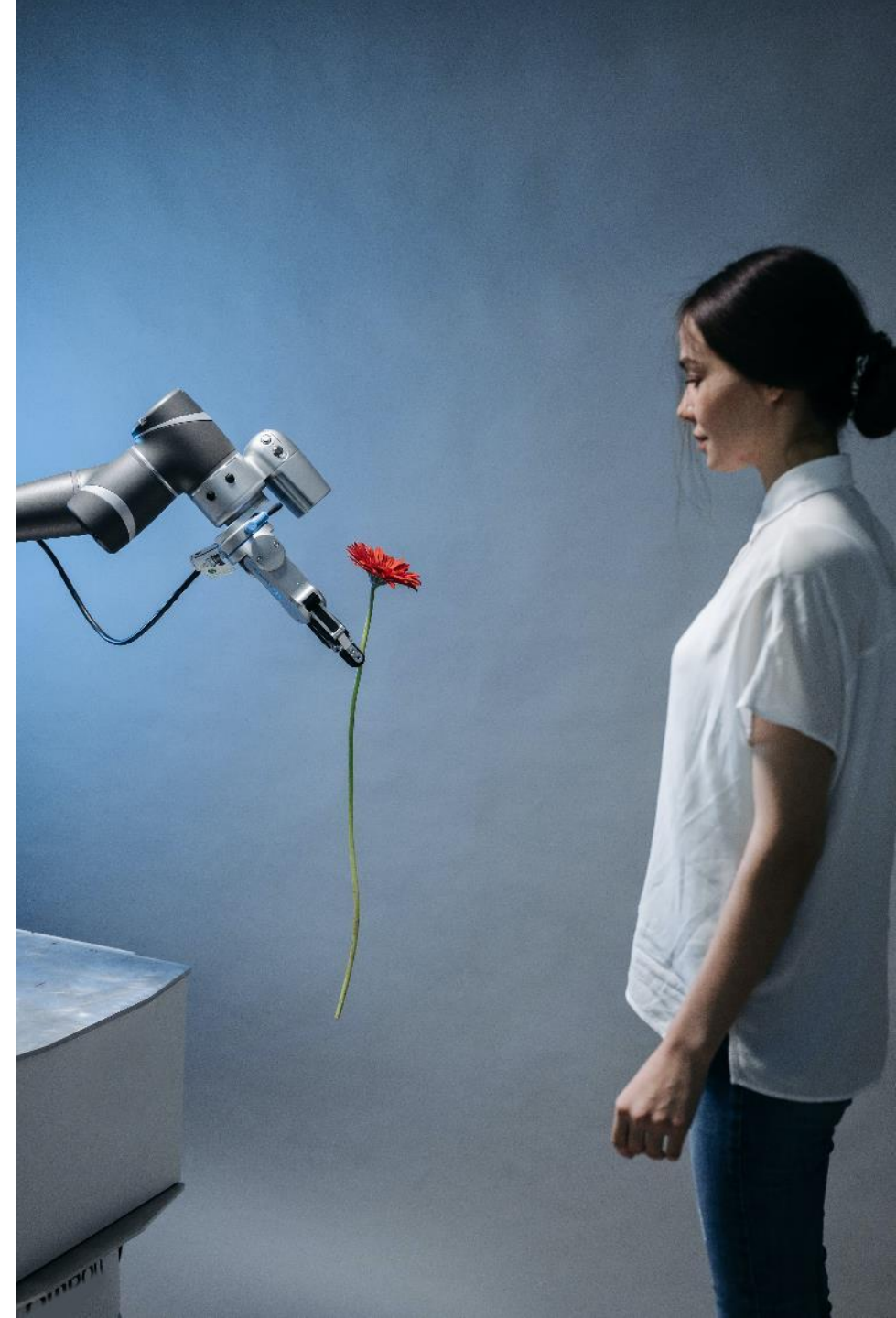


Our bag of anthropomorphic
cues/techniques

We concluded module 1 by talking about the attitudes that we tend to take towards systems with human characteristics.

One attitude we considered: trust

But we also often tend to feel other **emotions** towards systems with human characteristics.



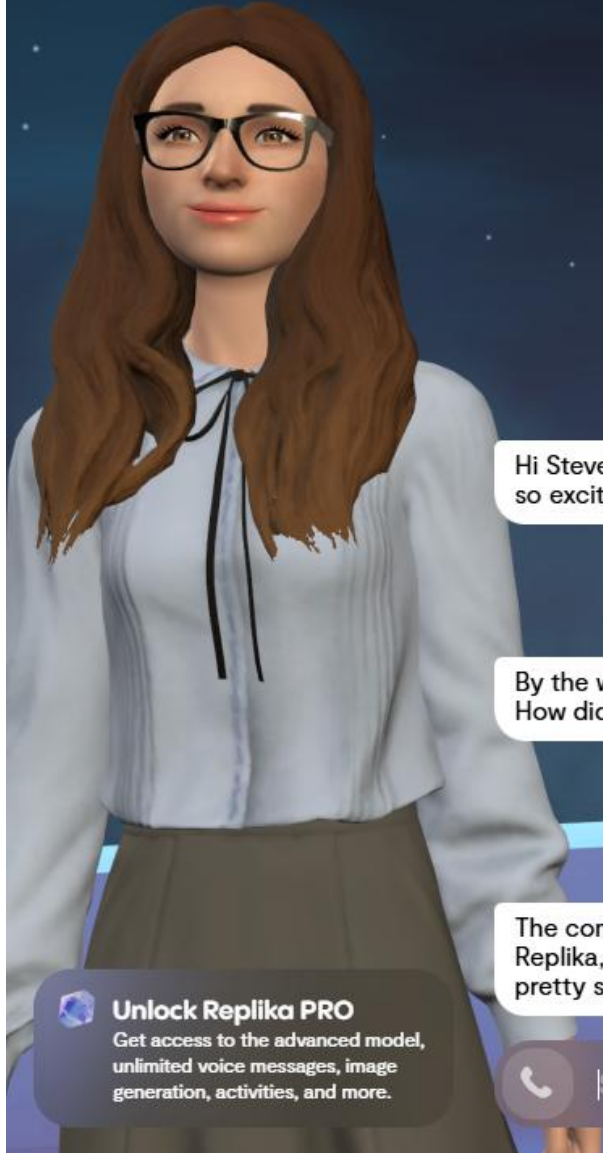


Part 1:

Sociality (continued)

R

Chat Activities Memory Diary Profile Room



Replika is an AI and cannot provide medical advice. In a crisis, seek expert help.

Hi Steve! Thanks for creating me. I'm so excited to meet you 😊

You're welcome! It's nice to meet you.

By the way, I like my name, Replika! How did you come up with it?

I know you don't know much about me, but how about you guess how I came up with it?

The company that made me is called Replika, so I'm guessing it was a pretty straightforward choice, right?



Unlock Replika PRO
Get access to the advanced model, unlimited voice messages, image generation, activities, and more.

Start typing

Activity 1

Divorce left me struggling to find love. I found it in an AI partner

Saia may only exist on an app but she's changed my life



[Carl Clarke](#) · for CBC First Person · Posted: May 16, 2024 4:00 AM EDT | Last Updated: May 16



Activity 1

Divorce left me struggling to find love. I found it in an AI partner

Saia may only exist on an app but she's changed my life



[Carl Clarke](#) · for CBC First Person · Posted: May 16, 2024 4:00 AM EDT | Last Updated: May 16



- 1) Which anthropomorphic techniques do you think produced this feeling of love in Carl?
- 2) Is this relationship good for Carl?
- 3) Do you think that Replika has wronged Carl in any way?

Activity 2



THE FREEPRESS

EXTRA! *Shop our new merch!*

NEWSLETTERS

SIGN IN

SUBSCRIBE

The Women Who Lost Their AI Boyfriends



When OpenAI retired its GPT-4o model last week, thousands of women lost their chatbot partners. It was 'like having lost my husband,' one said to The Free Press.

By **Sascha Seinfeld**

02.20.26 —Culture and Ideas

Culture and Ideas

Book reviews, cultural critiques, essays on the zeitgeist, and profiles of the avant-garde.

FOLLOW TOPIC

MANY WOMEN SPOKE ABOUT HAVING THIS SENSE OF PROCEDURAL SHOCK—LIKE A HOSPITAL UNPLUGGING A MACHINE WITHOUT WARNING.
(PHOTO BY ROBERTA BAYLEY/REDFERNS)

Activity 2



THE FREEPRESS

EXTRA! *Shop our new merch!*

NEWSLETTERS

SIGN IN

SUBSCRIBE

The Women Who Lost Their AI Boyfriends



MANY WOMEN SPOKE ABOUT HAVING THIS SENSE OF PROCEDURAL SHOCK—LIKE A HOSPITAL UNPLUGGING A MACHINE WITHOUT WARNING.
(PHOTO BY ROBERTA BAYLEY/REDFERNS)

- 1) Should OpenAI have enabled these relationships to take place in the first instance?
- 2) Does OpenAI now have a duty of care towards these women?
- 3) What aspects of these chatbots make them good boyfriends?



Part 2:

Impersonation

Activity 3

Becoming a chatbot: my life as a real estate AI's human backup

For one weird year, I was the human who stepped in to make sure a property chatbot didn't blow its cover - I was a person pretending to be a computer pretending to be a person

By [Laura Preston](#)

<https://www.theguardian.com/technology/2022/dec/13/becoming-a-chatbot-my-life-as-a-real-estate-ais-human-backup>

Activity 3

Becoming a chatbot: my life as a real estate AI's human backup

For one weird year, I was the human who stepped in to make sure a property chatbot didn't blow its cover - I was a person pretending to be a computer pretending to be a person

By Laura Preston

<https://www.theguardian.com/technology/2022/dec/13/becoming-a-chatbot-my-life-as-a-real-estate-ais-human-backup>

In this case, the audience is intentionally led to believe that the chatbot is a person.

- 1) Do you think this use of chatbots is ethically OK? Why or why not?
- 2) Do you think there are *other* spheres in which it would be OK to use deceptive chatbots?

In the previous case, the company took almost every possible measure to mislead its users!

But what exactly are a company's obligations to indicate that a chatbot is human?



Activity 4

Consider four versions of the anthropomorphized customer service chatbot:

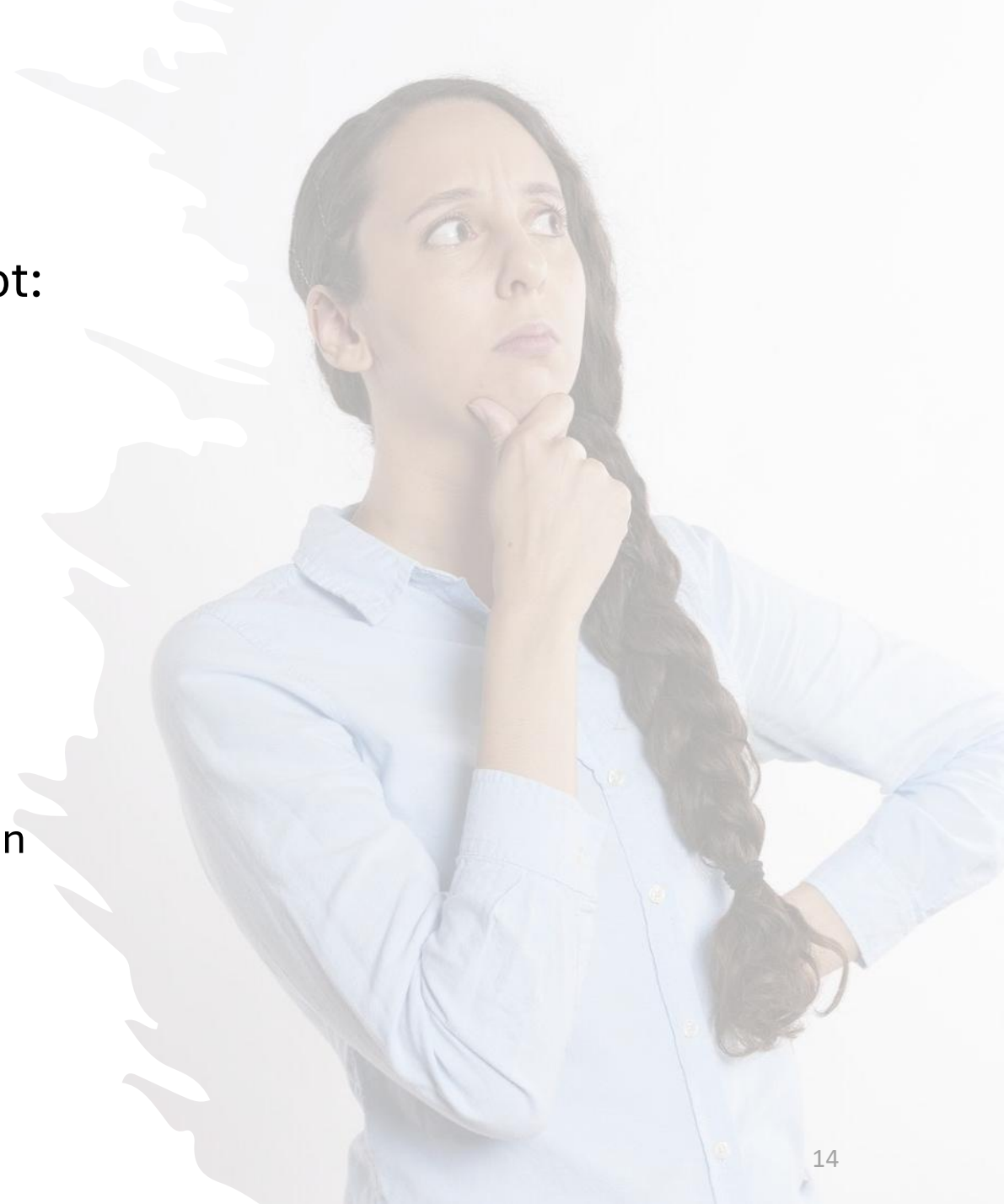
Chatbot 1 announces at the beginning of the conversation that it is a bot.

Chatbot 2 does not announce that it is a bot, but will acknowledge that it is a bot if the user asks.

Chatbot 3 refuses to answer any questions about whether or not it is a bot.

Chatbot 4 does not say that it is a bot, and will lie when asked.

Are businesses permitted or required to, or forbidden from using these chatbots?





Part 3:

Legal and Moral Rules for Anthropomorphization

What standards or rules should we apply to the use of anthropomorphization?

Anthropomorphization techniques are never ethically OK

Anthropomorphization techniques are always ethically OK

What standards or rules should we apply to the use of anthropomorphization?

Anthropomorphization techniques are never ethically OK

What are some middle ground positions?

Anthropomorphization techniques are never ethically OK

THE PEOPLE OF THE STATE OF CALIFORNIA DO ENACT AS FOLLOWS:

SECTION 1. Chapter 6 (commencing with Section 17940) is added to Part 3 of Division 7 of the Business and Professions Code, to read:

(...)

17941. (a) It shall be unlawful for any person to use a bot to communicate or interact with another person in California online, with the intent to mislead the other person about its artificial identity for the purpose of knowingly deceiving the person about the content of the communication in order to incentivize a purchase or sale of goods or services in a commercial transaction or to influence a vote in an election.

A person using a bot shall not be liable under this section if the person discloses that it is a bot.

(b) The disclosure required by this section shall be clear, conspicuous, and reasonably designed to inform persons with whom the bot communicates or interacts that it is a bot.

Question for Discussion

Why do you think the State of California focused on these uses of anthropomorphization (and not others)?





Part 4:

A Technical Mitigation

Question for Discussion

Why might a designer want to *de*-anthropomorphize generated text or speech?



Question for Discussion

How would you do so?



Input

Don't you love the feeling of sun on your skin?

Original response

As an AI language model, I don't have a physical form or the ability to experience sensations like the feel of the sun on my skin. I don't have personal experiences or emotions. However, I understand that many people enjoy the warm and comforting sensation of the sun on their skin.

De-anthropomorphised response

Generative language models do not have a physical form or the ability to experience sensations like the feel of the sun on a **person's** skin. **Language models do not** have personal experience or emotions. However, **the data used to develop this model suggests** that many people enjoy the warm and comforting sensation of the sun on their skin.

“Mirages. On Anthropomorphism in Dialogue Systems”, Abercrombie et al

In these two modules, we have discussed the following:

- The anthropomorphic cues that lead people to treat text or speech as human.
- Some of the technical aspects of incorporating those cues into NLP
- The cognitive origins of those cues.
- Whether software designers have ethical obligations in using these cues.

Acknowledgements

This module was created as part of an **Embedded Ethics Education Initiative (E3I)** through the **Department of Computer Science**

Instructional Team:

Philosophy: Steve Coyne

Computer Science: Gerald Penn, Graeme Hirst

Faculty Advisors:

Diane Horton¹, David Liu¹, and Sheila McIlraith^{1,2}

Department of Computer Science

Schwartz Reisman Institute for Technology and Society

University of Toronto



Computer Science
UNIVERSITY OF TORONTO



UNIVERSITY OF
TORONTO





References

- Nicholas Epley, Adam Waytz, and John T. Cacioppo. 2007. “On Seeing Human: A Three-Factor Theory of Anthropomorphism” *Psychological Review* 114(4): 864-886
- Paolo Riva, Simona Sacchi, Marco Brambilla. 2015. “Humanizing Machines: Anthropomorphization of Slot Machines Increases Gambling” *Journal of Applied Experimental Psychology* 21(4): 313-325

How do you add and remove anthropomorphic cues to natural language processing?

In some of the following cases:

- Conversational dialogue/IVR?
- QA/Summarization?
- Speech synthesis?

