COMPILER DESIGN
Reinhard Wilhelm · Dieter Maurer
FUNDAMENTALS OF COMPUTING II
Abstraction,

Embedded Ethics:

CSC401: Anthropomorphization (Module 2)

Slang

Embodiment

Explicit claims of humanness

First-person pronouns

Disfluencies

Warmth
in voice

Our bag of anthropomorphic cues/techniques

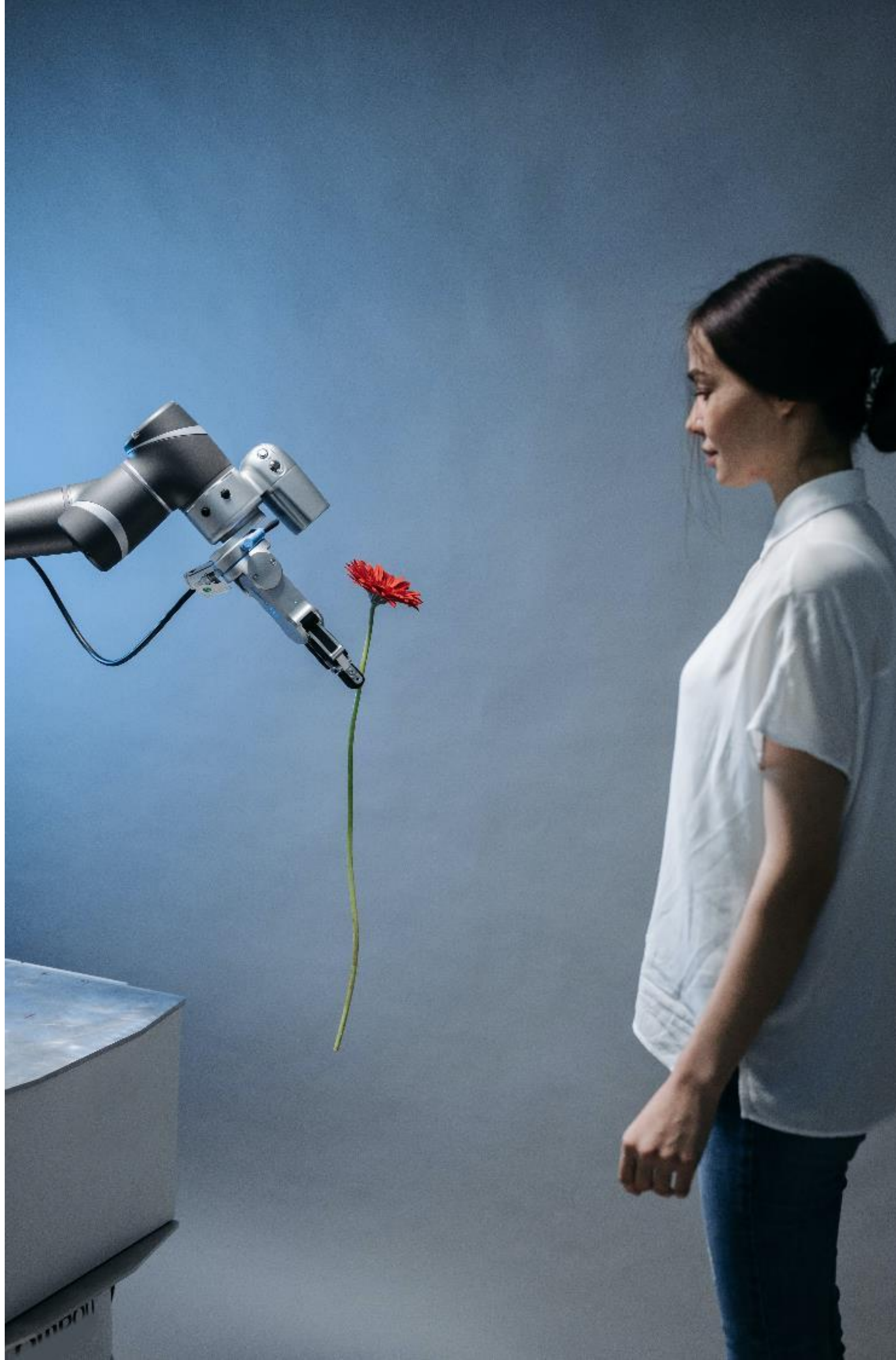
Part 1:

How to Anthropomorphize (CS-led, 5-10 minutes)

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?



Part 2:

The Ethics of Anthropomorphization

In the last module, we used effectance and sociality to talk about benefits and harms of anthropomorphization.

By themselves, benefits and harms don’t tell us about the ethics of anthropomorphize text or speech.

Have any of you talked to a human to do the following?

- Returning an item
- Cancelling a subscription
- Making a reservation

Introducing ChatGPT

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

It is likely that these tasks could be performed by a chatbot that uses many anthropomorphic techniques.

Activity 1

Consider four versions of the anthropomorphized customer service chatbot:

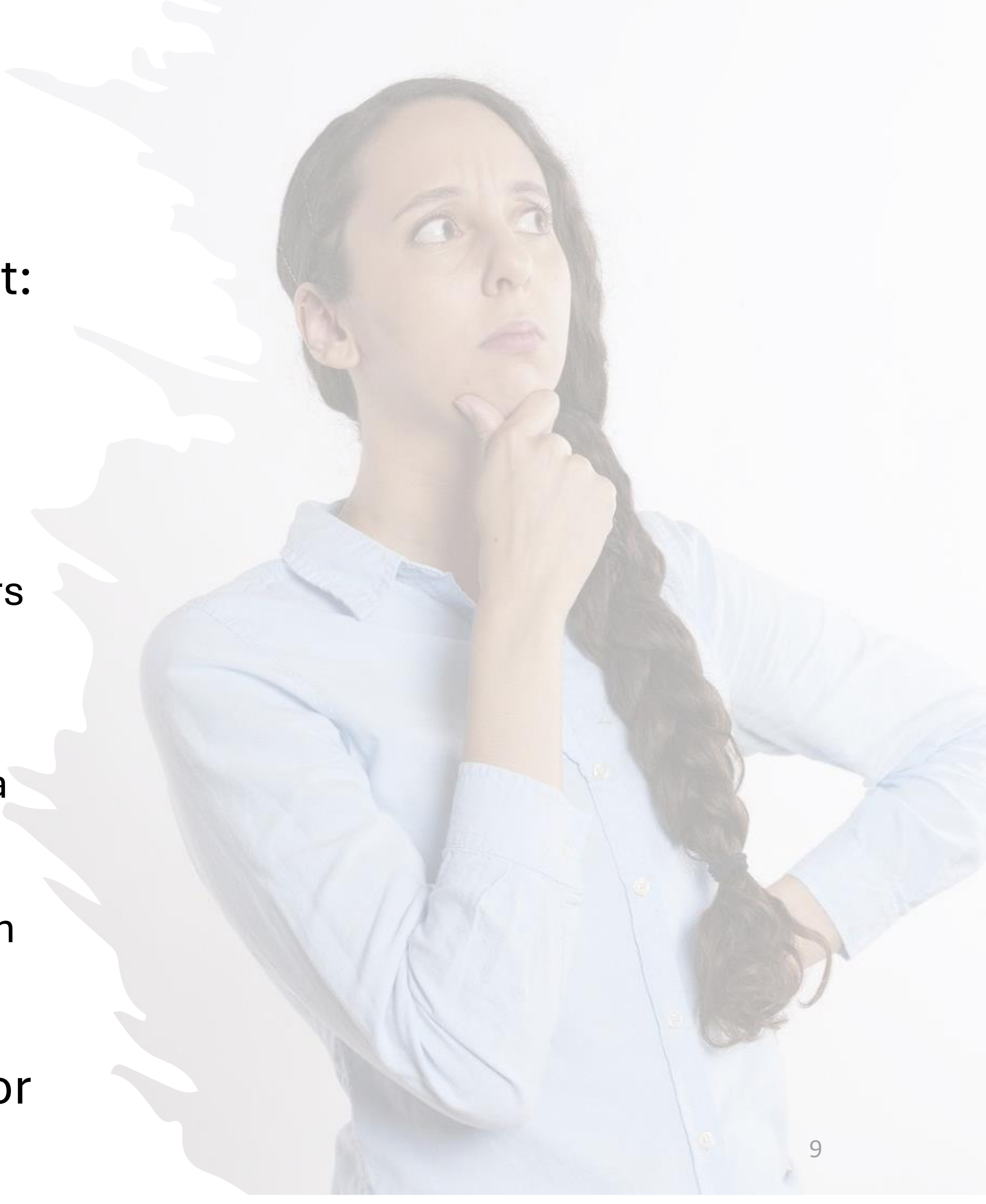
Chatbot 1 announces at the beginning of the conversation that it is a bot.

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

Chatbot 3 refuses to answer any questions about whether or not it is a bot. Many users believe that it is a human.

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

Would it be ethical for a business to use any or all of these chatbots?



Deception: the intentional attempt to produce a false belief in someone

Lying: uttering a sentence believed to be false with the intention of producing a false belief in someone

Slang

Embodiment

Explicit claims of humanness

First-person pronouns

Disfluencies

Warmth
in voice

Assume that the designer of an anthropomorphic system had the intention of deceiving the user that it was human. Would any of these count as lying?

Slang

Embodiment

Explicit claims of humanness

First-person pronouns

Disfluencies

Warmth
in voice

Are any of these techniques more or less deceptive than others?

Question for Discussion

One common defense of deception and lying is that people are not entitled to know certain information – e.g. it is maybe OK to lie if someone asks you a personal question.

Would you be entitled to know that a chatbot is human? Why?

Part 3:

Beyond Deception

Question for Discussion

If the user is not deceived by the use of anthropomorphization techniques, can the use of those techniques still treat them wrongly?

Example: anthropomorphized slot machines (Riva, Sacchi and Brambilla, 2015)

This is probably wrong! But why is it wrong?

Another example: Replika

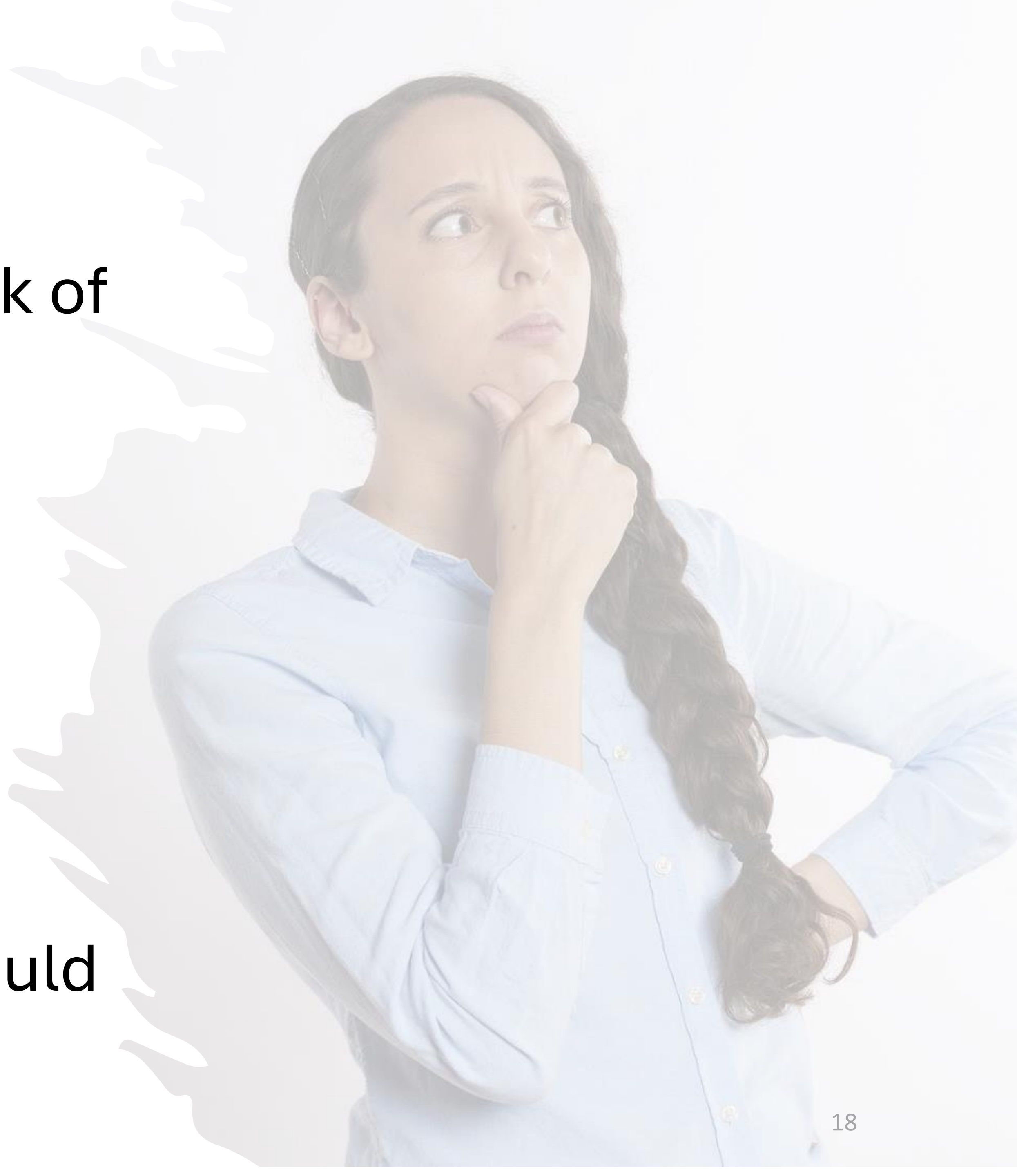
[Showcase Replika]

Activity 2

Let's assume that there is no risk of deception.

Should the creators of the customer service chatbot and Replika maximize their usage of anthropomorphic cues?
(voice cues, pronouns, etc)

If not, why do you think they should hold back, and for which ones?



Question for Group Discussion

How is the Replika case different from the slot machine case?

Some reasons to think actions are wrong have to do with their consequences (harms and benefits).

Some reasons to think actions are wrong have to do with the nature of those actions themselves (e.g. that they involve manipulation, deception or exploitation)

Part 5:

Legal and Moral Rules for Anthropomorphization (PHL and CS led, 10 minutes)

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

Middle ground positions?

Anthropomorphization techniques are never ethically OK

Question for Discussion

What are some middle ground positions between anthropomorphization being always OK and it being never OK?

Question for Discussion

What sorts of barriers are there to adopting these ethical rules as laws?

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.

In this module, we have discussed the following:

- The anthropomorphic cues that lead people to treat text or speech as human.
- 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: Graeme Hirst, Gerald Penn

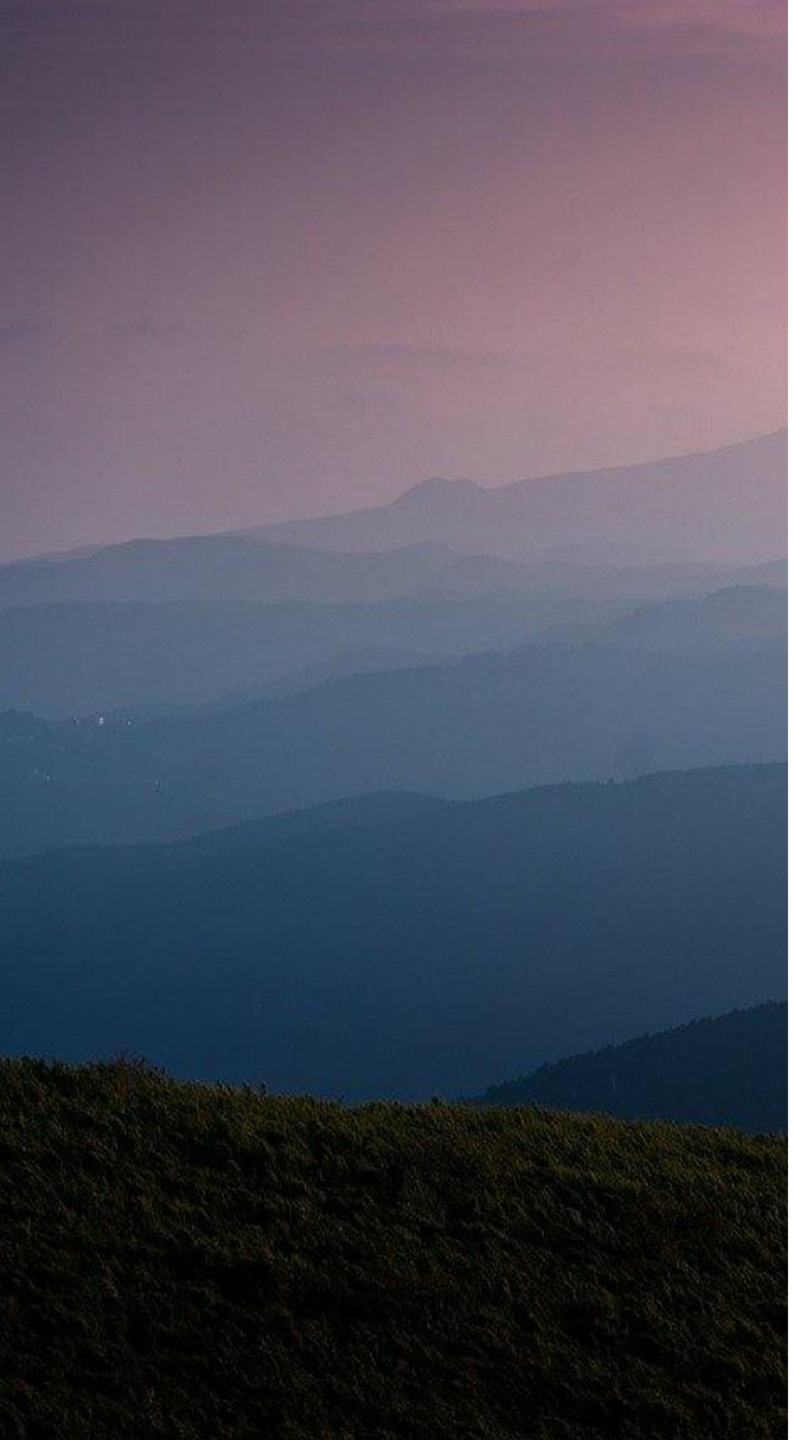
Faculty Advisors:

Diane Horton[1], David Liu[1], and Sheila McIlraith[1,2]

Department of Computer Science

Schwartz Reisman Institute for Technology and Society

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