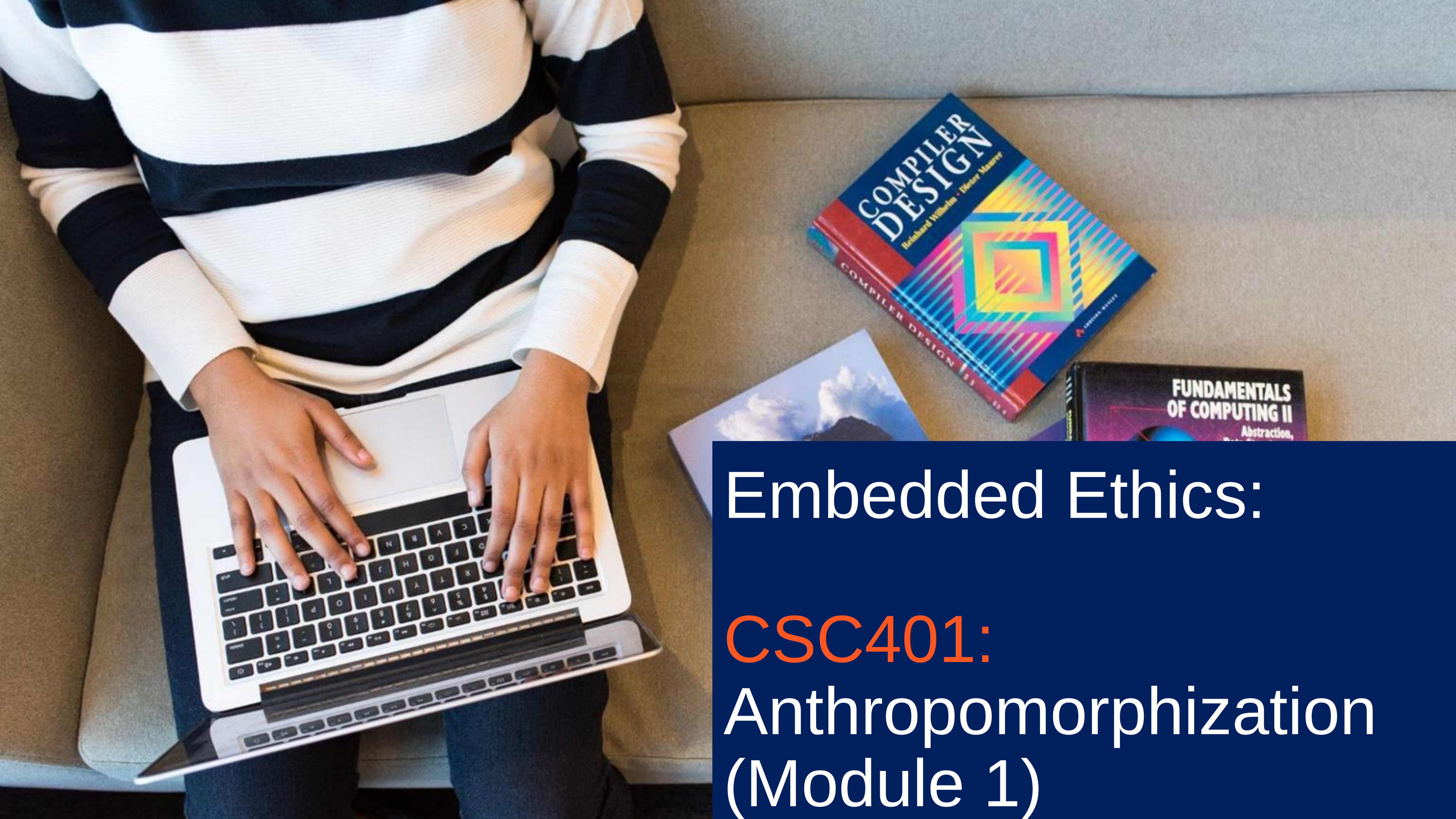

COMPILER DESIGN
Reinhard Wilhelm · Dieter Maurer
COMPILER DESIGN
FUNDAMENTALS OF COMPUTING II
Abstraction,

Embedded Ethics:

CSC401: Anthropomorphization (Module 1)

Part 1:

Introduction

Welcome to Embedded Ethics!

This embedded ethics module is a collaboration between philosophy and computer science.

A goal of embedded ethics is to give you the skills and incentive to recognize ethical issues that arise in the development of software and to integrate ethical considerations into your work and research as computer scientists.

Welcome to Embedded Ethics!

This module contains discussion questions and group activities, but also feel free to ask questions or make comments by raising your hand.

In CSC401 you have been working with computational and statistical models of language.

These models are a building block for synthesizing new text or speech:

- LLMs
- Language translation
- Speech synthesis

A system that has features that make it appear to be human is anthropomorphic.

Anthrōpos = human
Morphē = form

When creating programs that produce text or speech, you have a choice:

- Should you anthropomorphize the created text or speech?
- That is, should you make it seem human?

Today, we are going to help you understand:

- How to incorporate anthropomorphic techniques into natural language systems.
- Some of the benefits and risks of anthropomorphizing systems

Part 2:

Anthropomorphic Cues

FIGURE 01 + OPENAI
SPEECH-TO-SPEECH REASONING

What makes an auditory clip seem more or less human-sounding?

Voice

- Accent or tone: some ways of speaking register as more “human” sounding
  - Warmth
  - Breathiness

Voice

- Disfluencies: elements that break the flow of speech
  - Pauses
  - Hesitations
  - Filler words (e.g. “um”)

Question for Discussion

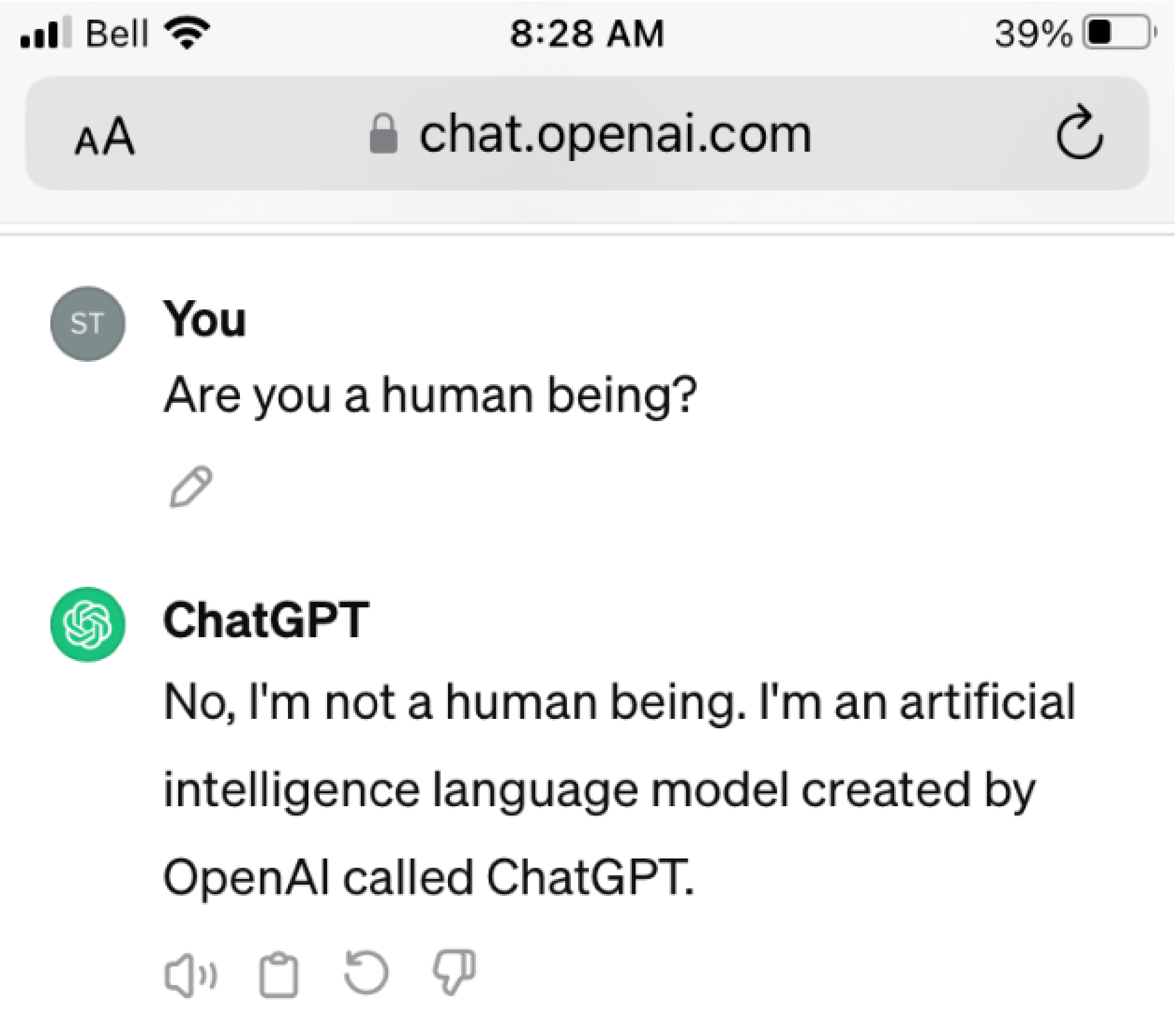
What are some elements of written text that make generated text seem to be authored by a human?

Register and Style

- Slang or informal speech

Content

- How the system answers questions about itself
- The information that the system spontaneously offers about itself



You

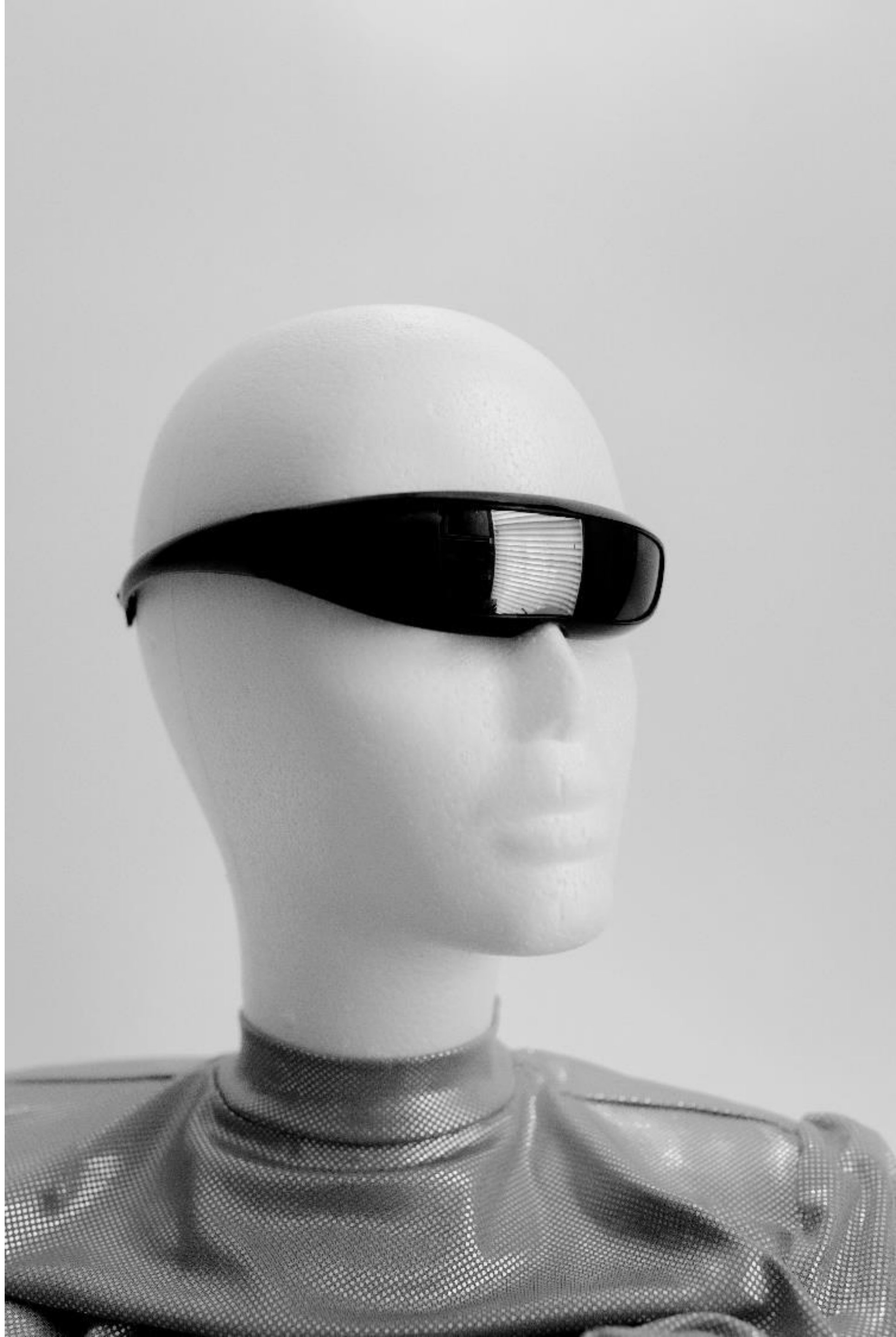
Are you a human being?

ChatGPT

No, I'm not a human being. I'm an artificial intelligence language model created by OpenAI called ChatGPT.

Content

- Use of first-person “I”, “me”, etc
- Implicit or explicit claims of having certain capabilities (e.g. having a personality, emotions)
- Explicit claims of being human (e.g. “I am human.”)



Embodiment

- We treat systems with human faces and bodies as more human-like

Slang

Embodiment

Explicit claims of humanness

First-person pronouns

Disfluencies

Warmth
in voice

Our bag of anthropomorphic
cues/techniques

Question for Discussion

What are some cases where a designer might want to anthropomorphize generated text or speech?

(Recall the systems in CSC401: LLMs, speech synthesis..)

Part 3:

Effectance

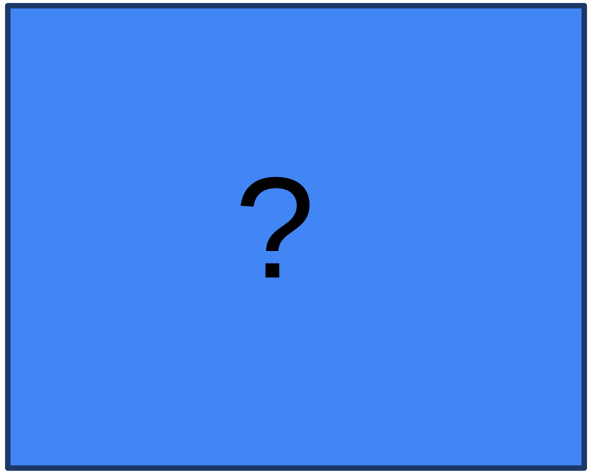
Epley et al., 2007:

As humans, we naturally tend to treat things as human, even when we know they are not human.

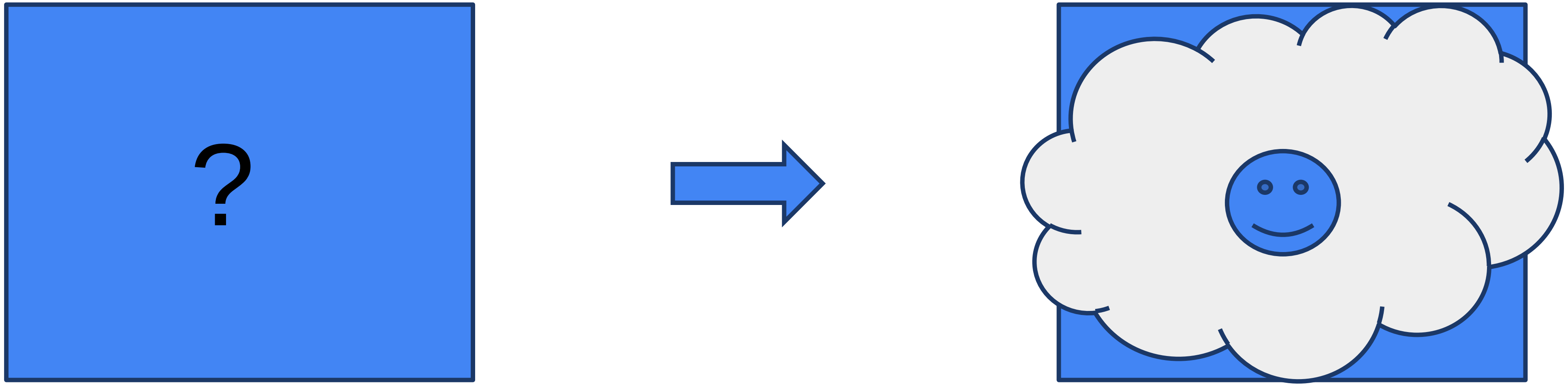
Our tendency to do so has two origins:

- Effectance
- Sociality

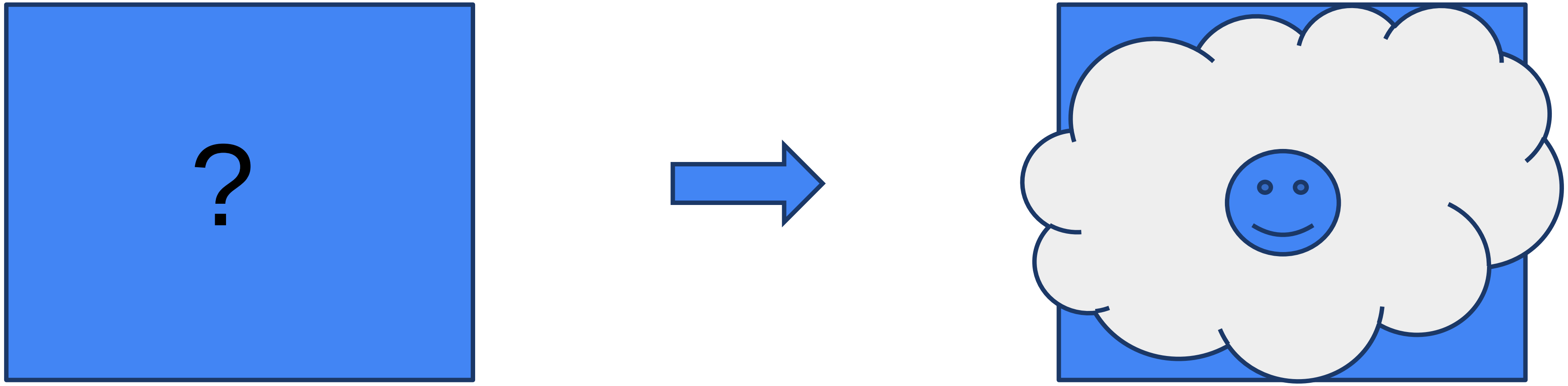
Sometimes we encounter systems that we don’t fully understand: we don’t know how they will behave.



We model these unknown systems: we treat them like things that we better understand. This helps us make predictions about how they will behave and how we should interact with them.

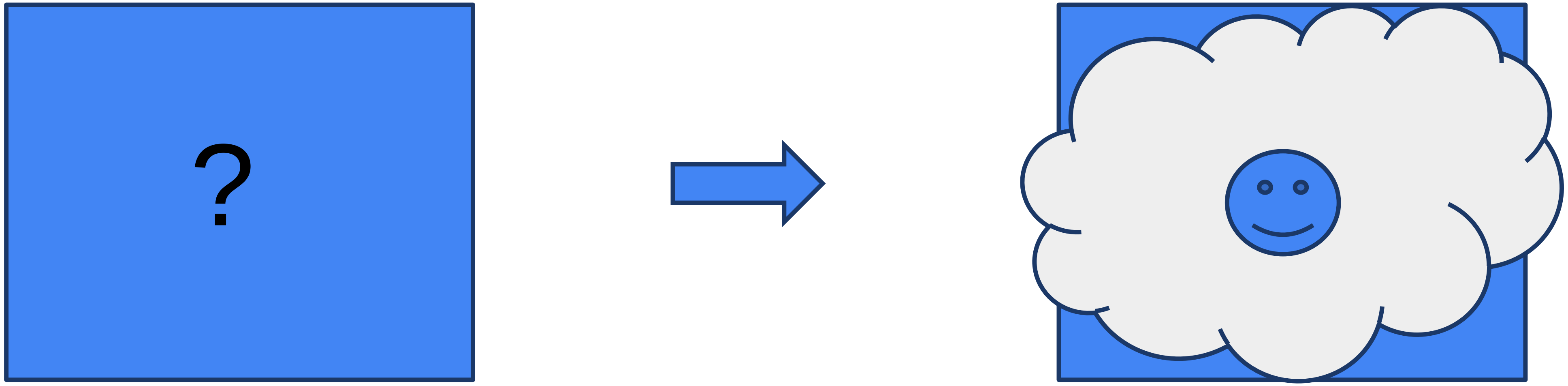


We model these unknown systems: we treat them like things that we better understand. This helps us make predictions about how they will behave and how we should interact with them.

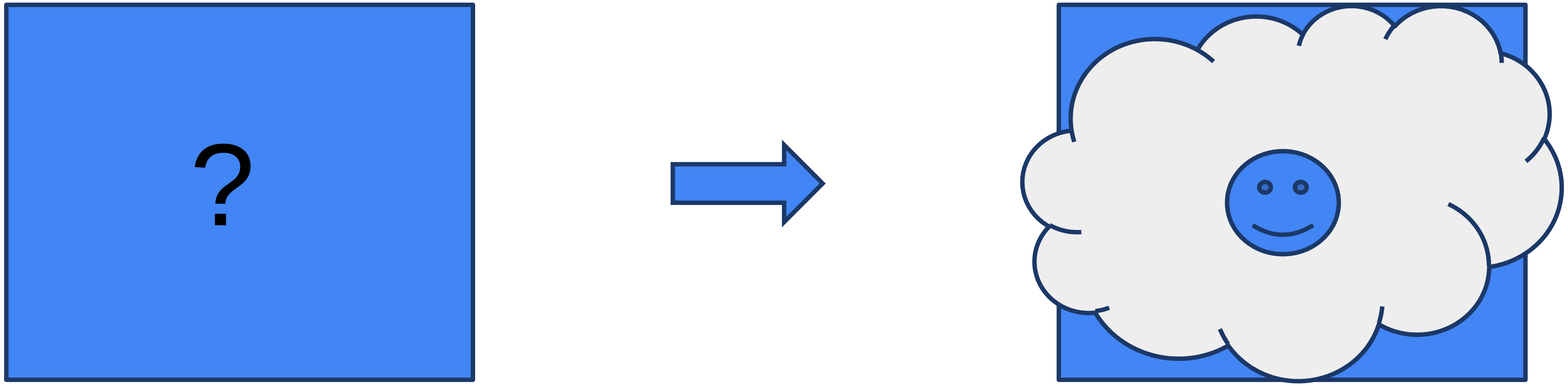


To be clear: when we model unknown systems, we usually know that the unknown system is not what we are modelling it as!

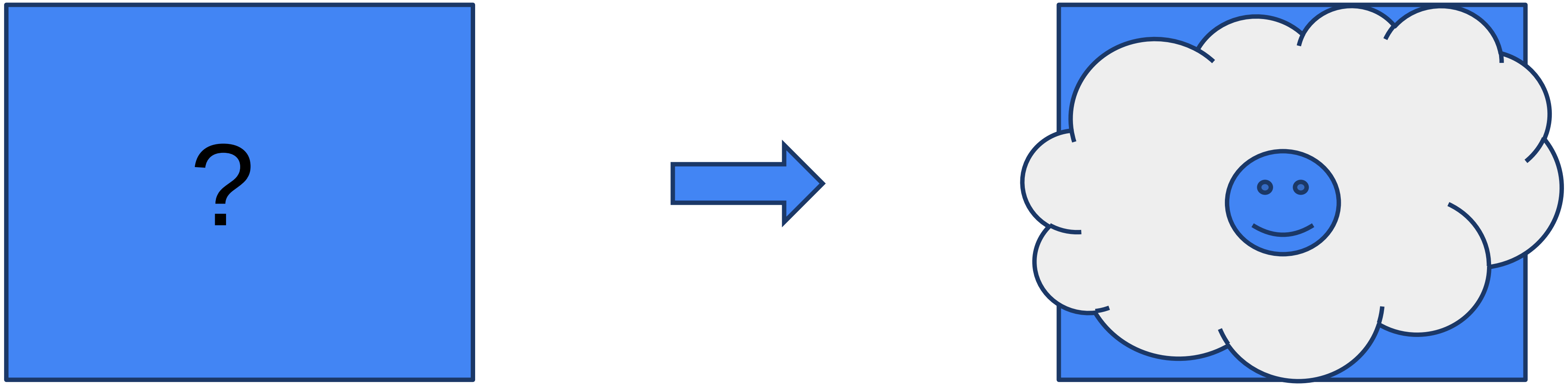
This applies to the case of pets, too!



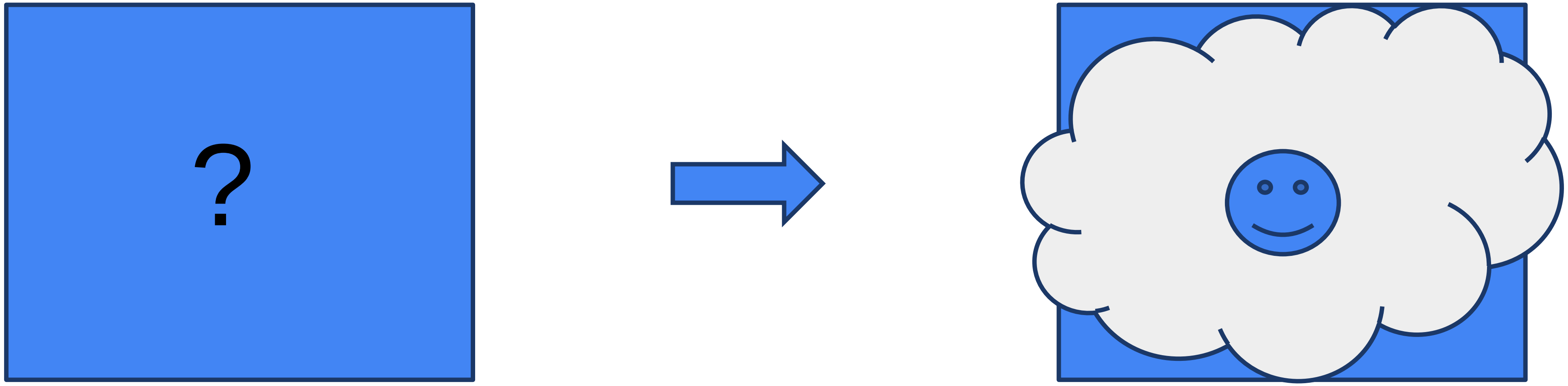
If we model a system as a human being, we tend to attribute some uniquely human abilities to it.



Question: What are some abilities that humans uniquely have?

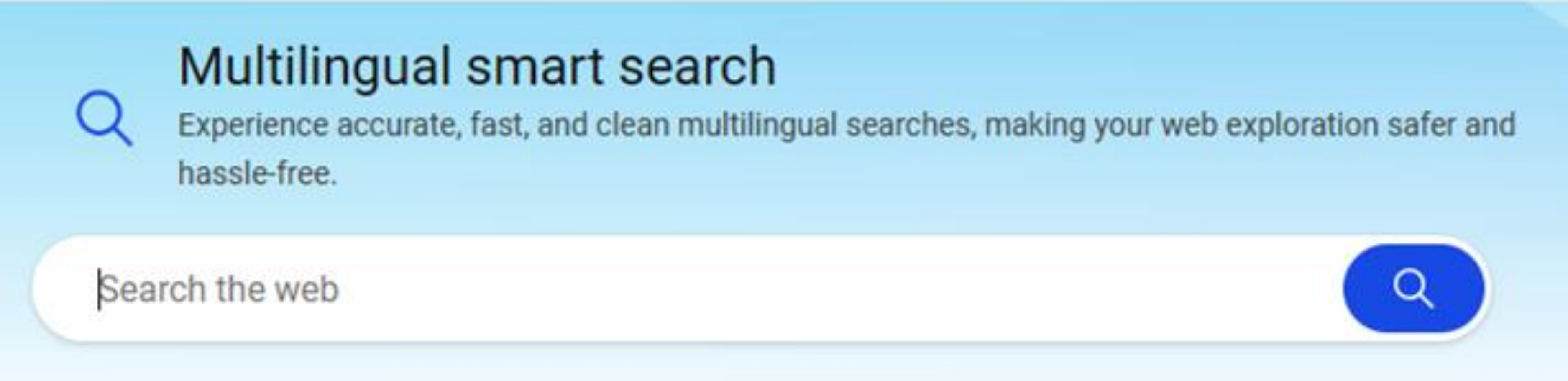


Memory
Emotions
Rational inference
Ability to understand questions



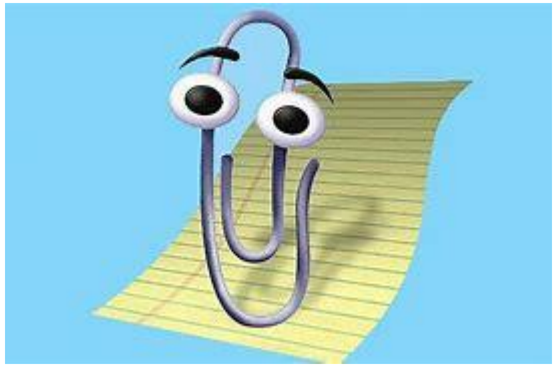
When the modelled system actually has these capabilities, this can **assist** the user in using the system.

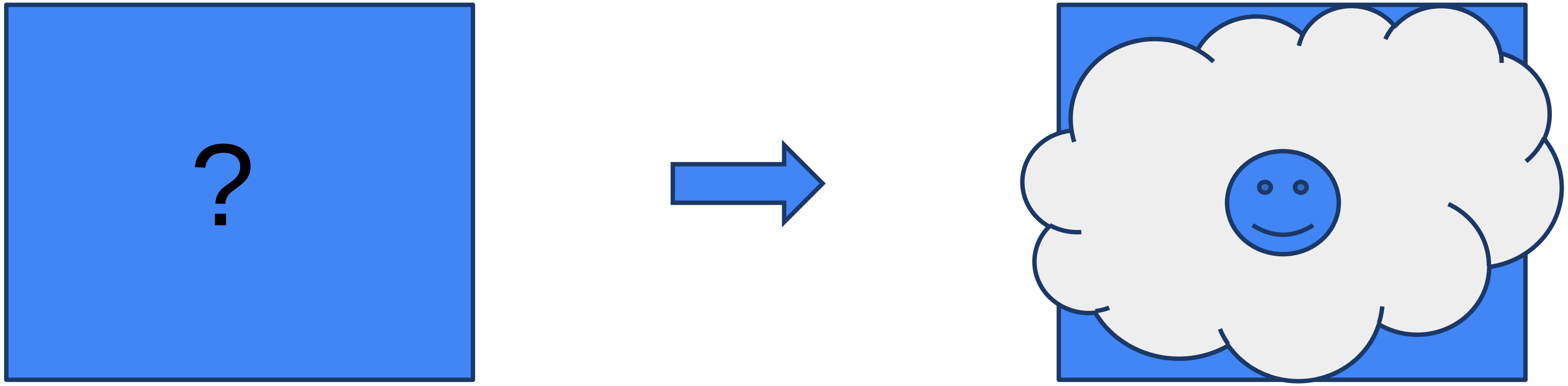
Memory
Emotions
Rational inference
Ability to understand questions



Multilingual smart search

Experience accurate, fast, and clean multilingual searches, making your web exploration safer and hassle-free.



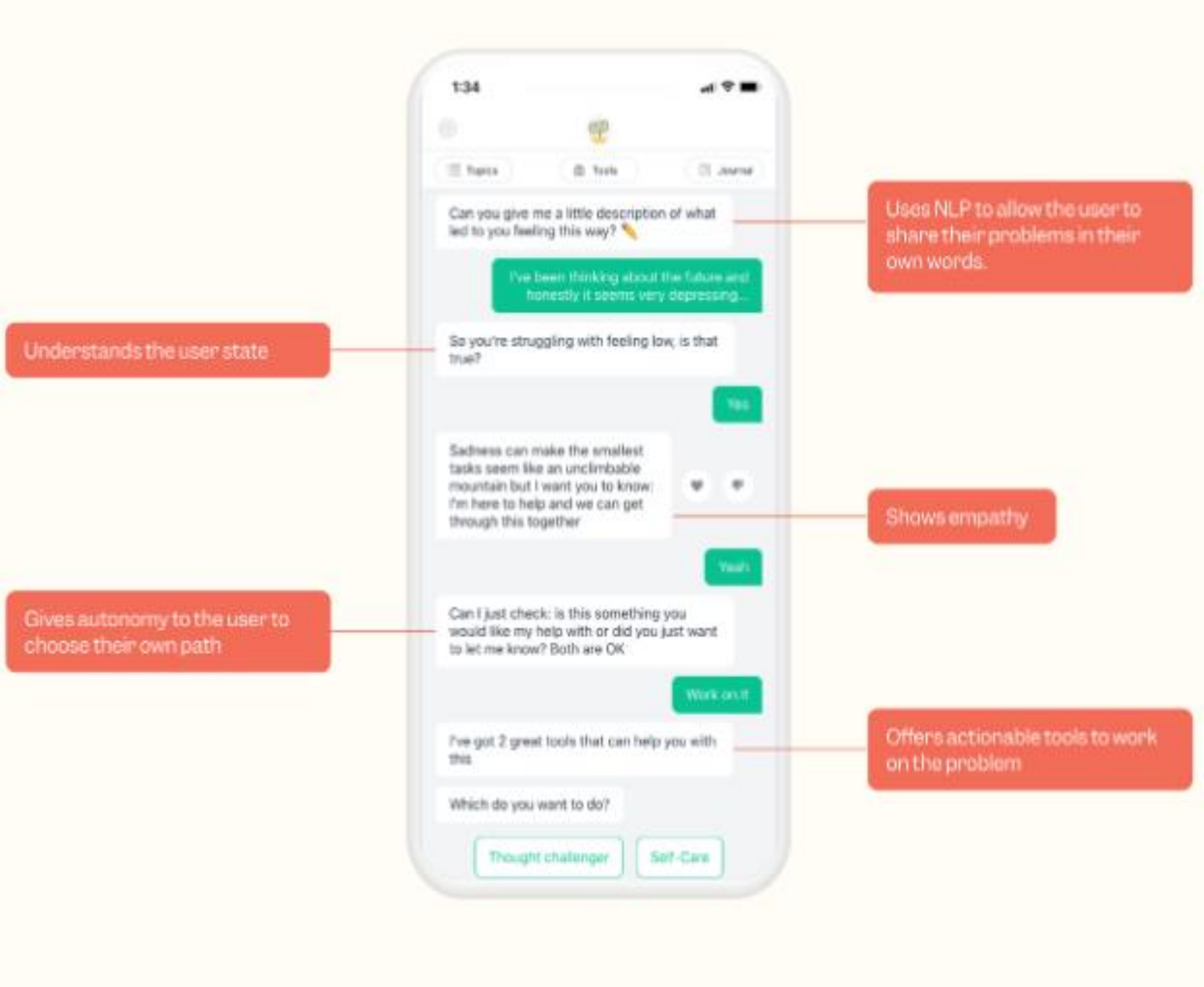


Memory
Emotions
Rational inference
Ability to ask questions

When the modelled system does not actually have these capabilities, this can **harm** the user.

Group Activity: Therapy bots (Woebothealth.com)

1. What capabilities might a user assign to the chatbot that they wouldn't assign to google search?
2. Which of those capabilities would the chatbot likely have, and how could that assist the user?
3. Which of those capabilities would the chatbot not likely have, and how could that harm the user?



Can you give me a little description of what led to you feeling this way? ✏️

I've been thinking about the future and honestly it seems very depressing...

So you're struggling with feeling low, is that true?

Yes

Sadness can make the smallest tasks seem like an unclimbable mountain but I want you to know: I'm here to help and we can get through this together

Yeah

Can I just check: is this something you would like my help with or did you just want to let me know? Both are OK

Work on it

I've got 2 great tools that can help you with this

Which do you want to do?

Uses NLP to allow the user to share their problems in their own words.

Understands the user state

Shows empathy

Gives autonomy to the user to choose their own path

Offers actionable tools to work on the problem

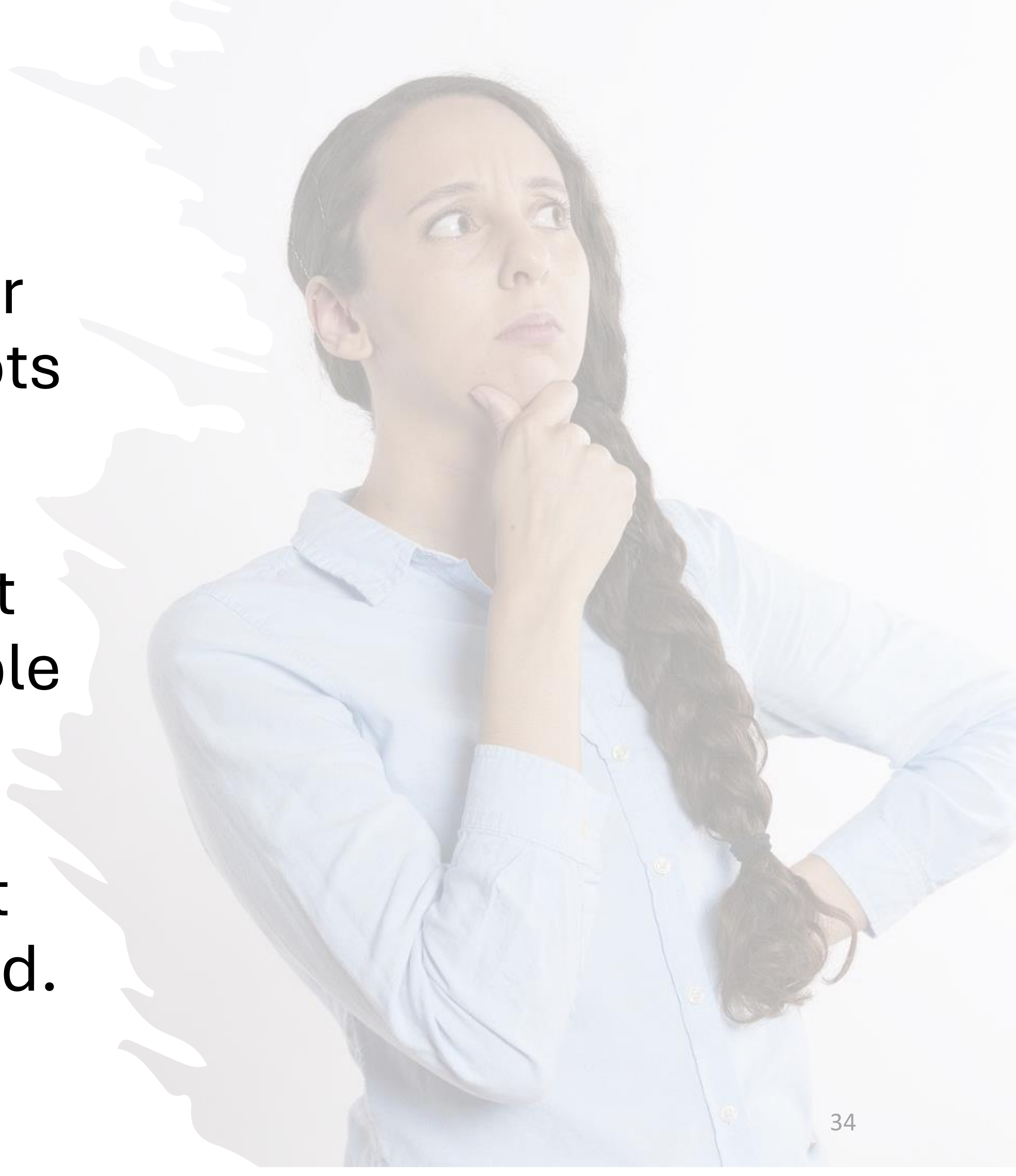
Human capability a user might attribute to an anthropomorphized therapy chatbot	Does the chatbot have it?	Potential harm or benefit

Activity 2

To what degree should customer service chatbots and therapy bots use anthropomorphic cues?

Assign it a number from 0 (not at all) to 5 (to the maximum possible extent).

Assume that there is no risk that the user will actually be deceived.



Part 4:

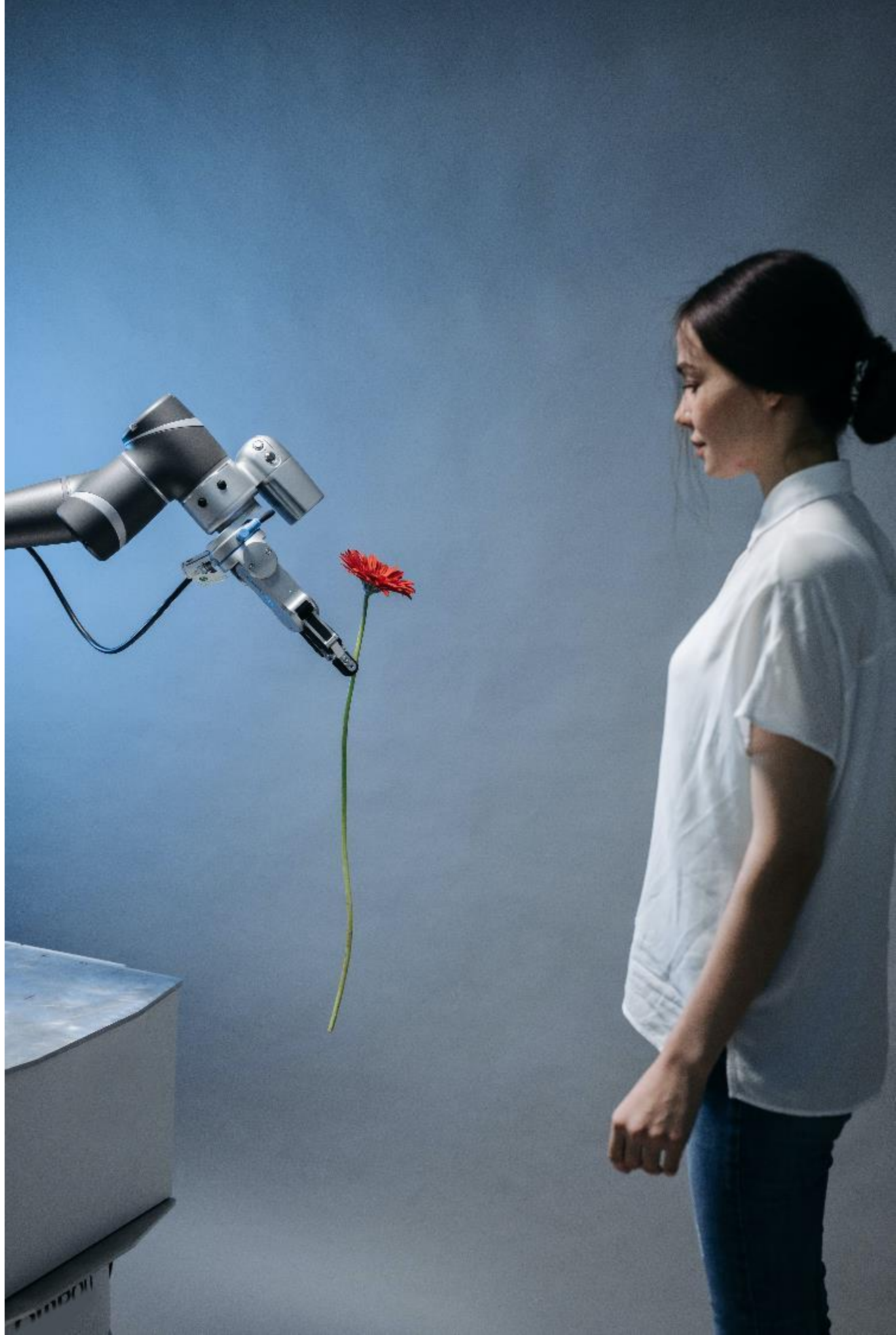
Sociality

Activity 2

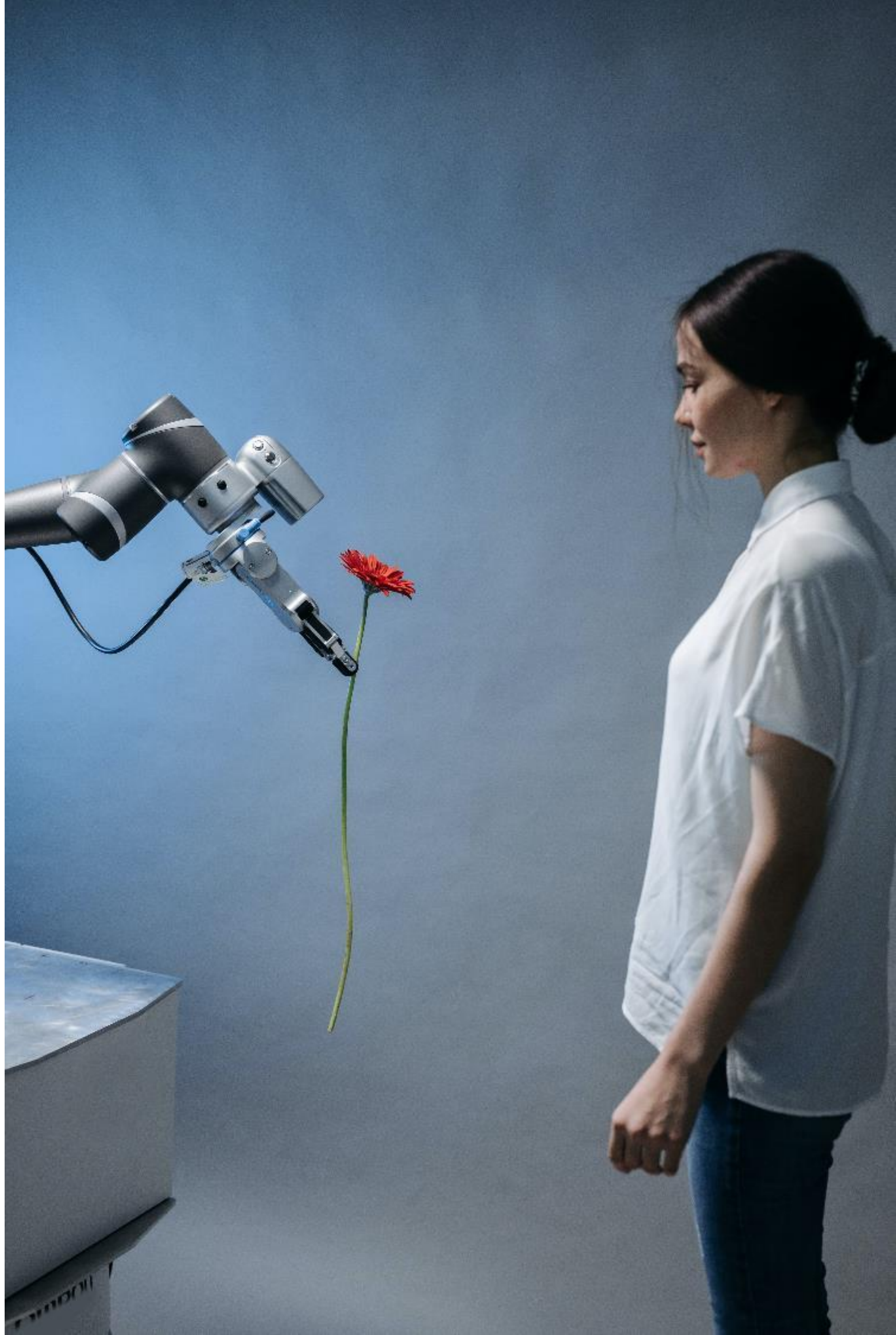
So far we have been talking about what assumptions we tend to make about systems with human characteristics.

What attitudes do we tend to take towards them?

Sociality: we mentally construct systems as humanlike to fulfil a need for social connection. (Epley, 2)



Users are more inclined to trust systems with which they feel a social connection.



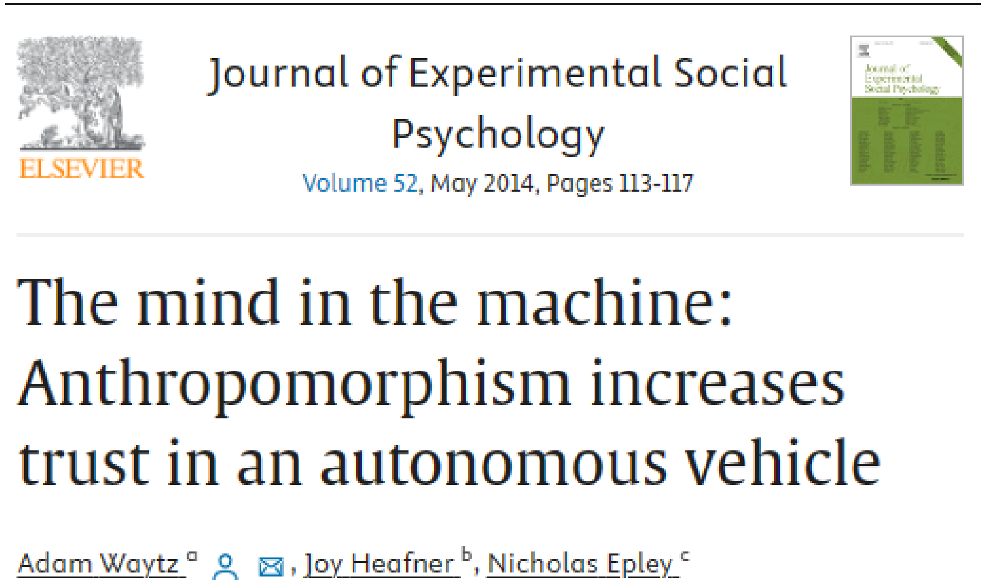
Question for Discussion

How is our behaviour different towards people we trust than people we don’t trust?

Question for Discussion

What is a way that the trust generated by anthropomorphization might be used to benefit the user?

, ,



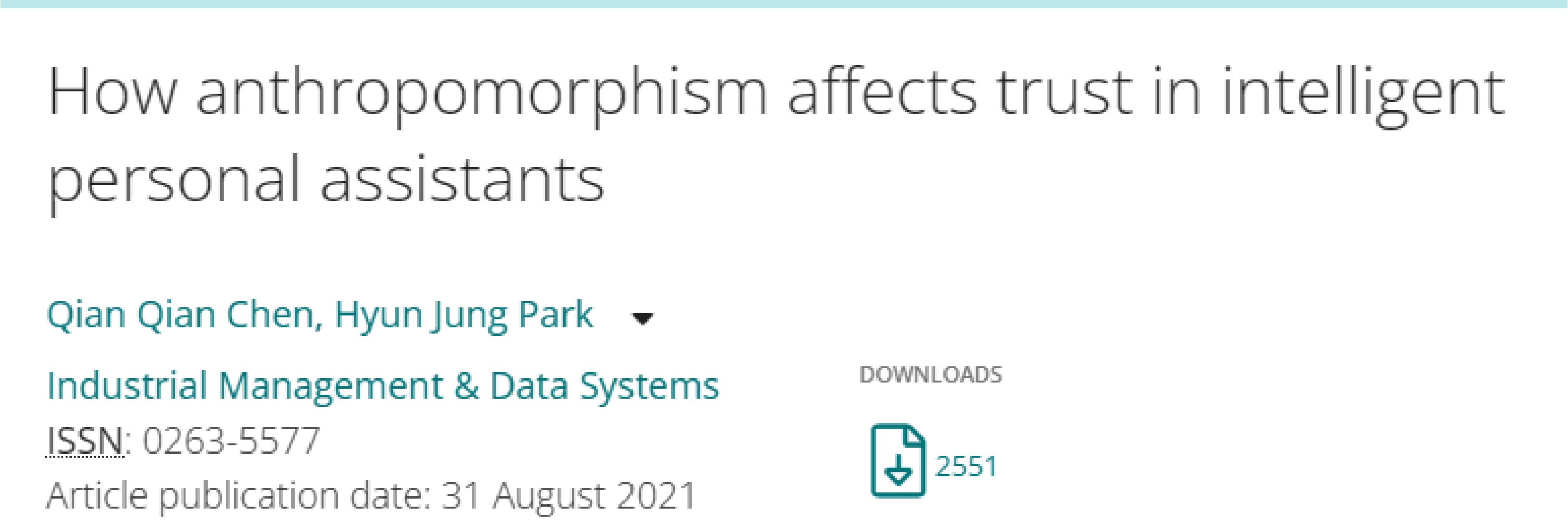
ELSEVIER

Journal of Experimental Social Psychology

Volume 52, May 2014, Pages 113-117

The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle

Adam Waytz [a], Joy Heafner [b], Nicholas Epley [c]



How anthropomorphism affects trust in intelligent personal assistants

Qian Qian Chen, Hyun Jung Park

Industrial Management & Data Systems

ISSN: 0263-5577

Article publication date: 31 August 2021

Question for Discussion

What is a way that the trust generated by anthropomorphization might be used to harm the user?

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

No one is deceived by the anthropomorphization used by these machines.

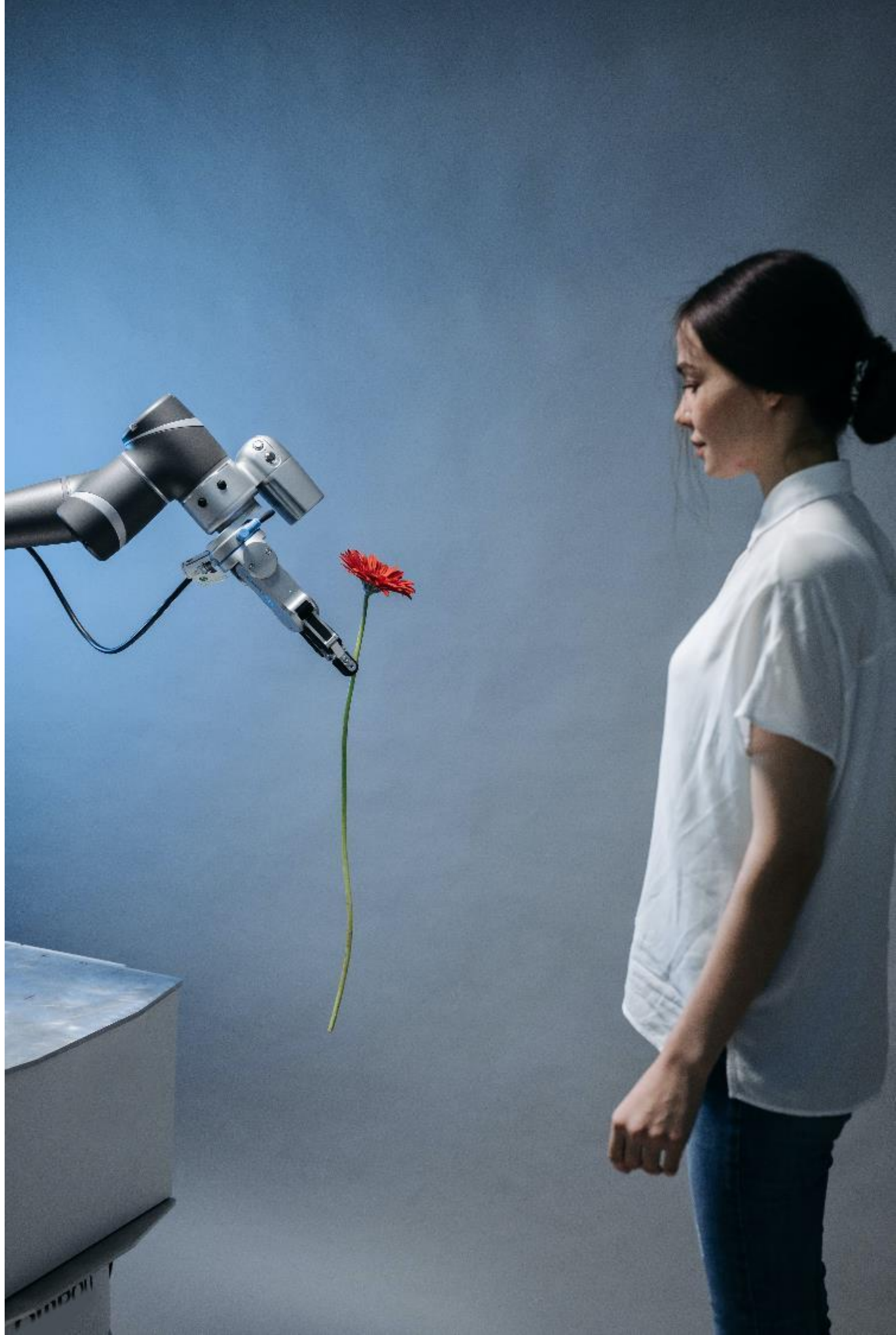
What do you think is going on here?

In module 2….

We will talk about more controversial cases of anthropomorphization:

- Anthropomorphization and intimacy
- The use of anthropomorphization to intentionally impersonate human beings
- Moral and legal rules that should govern anthropomorphism

See you then!



Acknowledgements

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

Instructional Team:

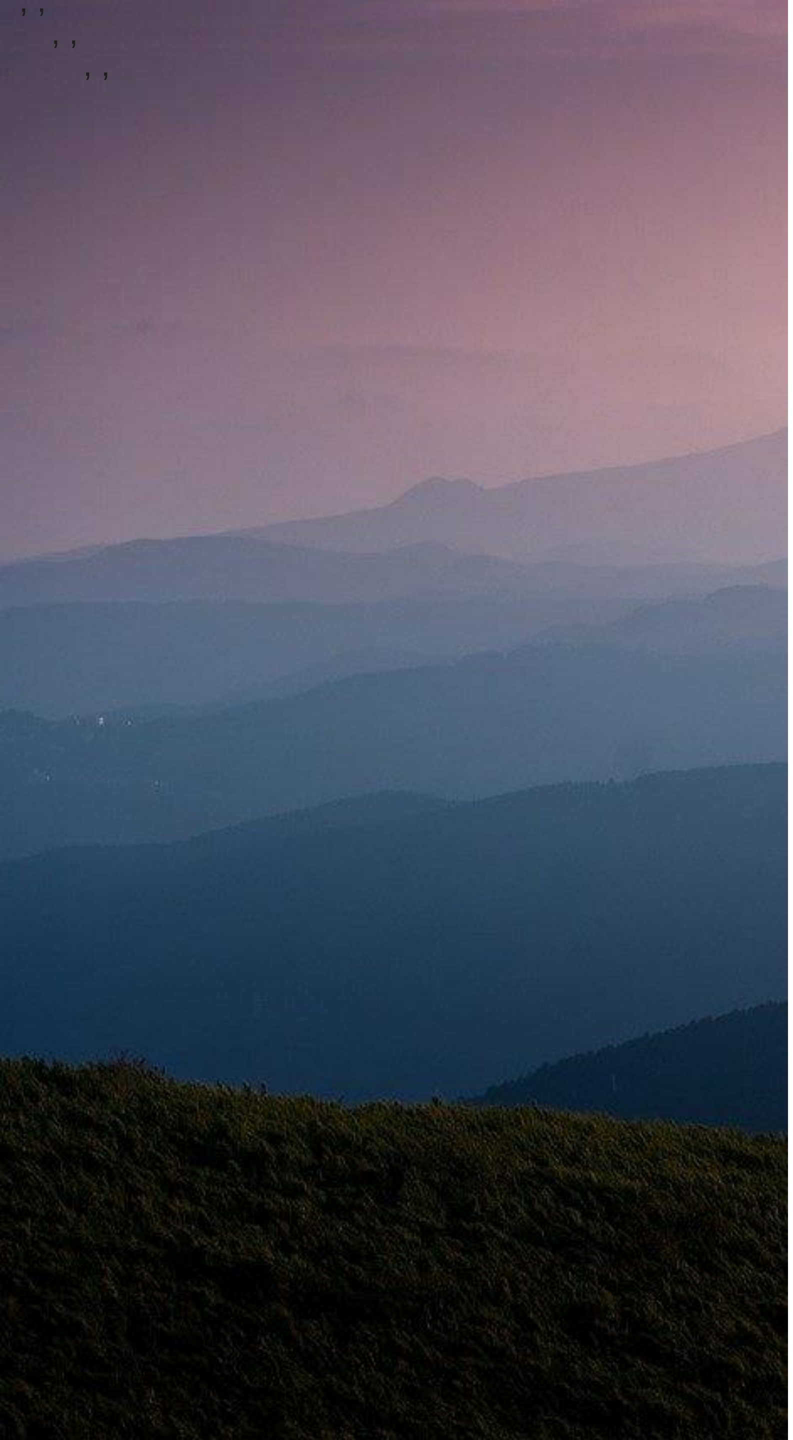
Philosophy: Steve Coyne

Computer Science: Gerald Penn, Graeme Hirst

Faculty Advisors:

Diane 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

- Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar and Zeerak Talat. 2023. “Mirages: On Anthropomorphism in Dialogue Systems”
- Nicholas Epley, Adam Waytz, and John T. Cacioppo. 2007. “On Seeing Human: A Three-Factor Theory of Anthropomorphism” *Psychological Review* 114(4): 864-886
- Adam Waytz, Joy Heafner, Nicholas Epley. 2014. “The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle” *Journal of Experimental Social Psychology* 52: 113-7