Theories of AI

Rick Nouwen

Utrecht
University

Road map

Conversational agents based on large language models

1. Linguistic pragmatics $\sim$ the AI boom
   - machine perspective
   - human perspective
2. Linguistic pragmatics $\sim$ Human-AI interaction
3. Some early (non)results

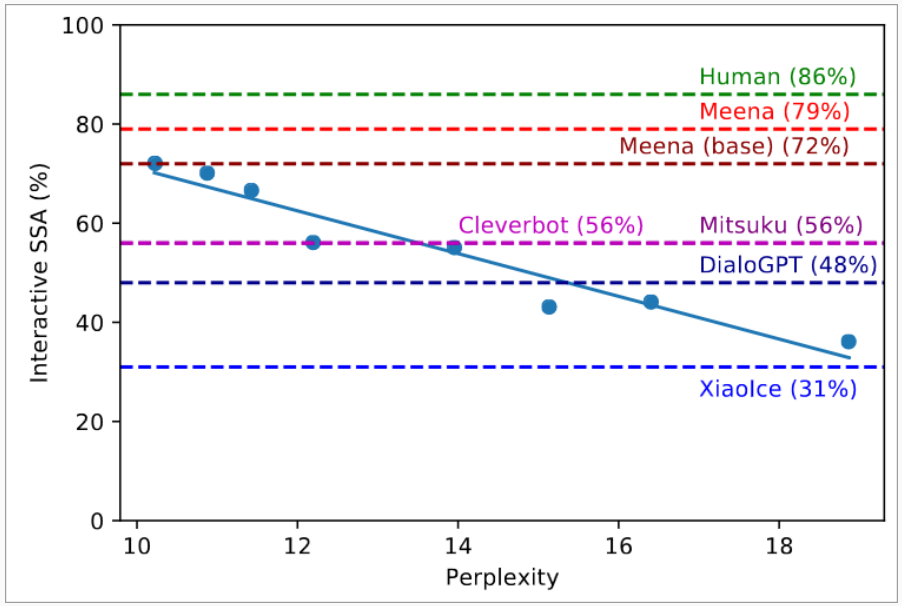
The role of pragmatics in the AI boom

- Pragmatics as a benchmark for human-likeness
- Pragmatics as domain to learn about emergent behaviour in machine learning
- Pragmatic challenges for AI

Pragmatic behaviour and human-likeness

Adiwardana et al. (2020). Towards a human-like open-domain Chatbot, arxiv 2001.09977

- Chatbot *Meena*, 40B word multi-turn dialogs
- Crowd-sourced measure of human-likeness: SSA
- For each response:
  - Sensibleness: is the response truthful and relevant?
  - Specificity: is the response specific to the context?



Emergent pragmatic understanding

Ruis et al. (2023). The Goldilocks of Pragmatic Understanding: Fine-Tuning Strategy Matters for Implicature Resolution by LLMs. arXiv:2210.14986

Esther asked "Can you come to my party on Friday?" and Juan responded "I have to work", which means `YES/NO`

base models	
BERT	54.8%
GPT3	57%
instruction tuned	
ChatGPT	71.1%
GPT-4	81.8%
human	
average	86.2%
best	92%

- limited pragmatic understanding arises in pre-training
- boost of pragmatic understanding on the basis of *instruction tuning*

Utrecht University

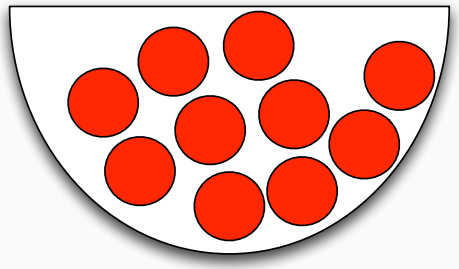
Pragmatics and (un)groundedness

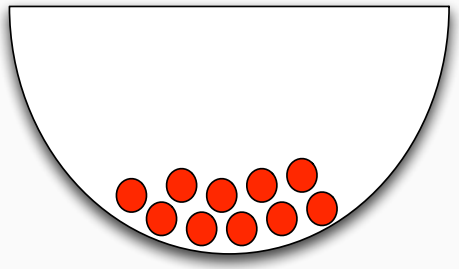
- Use of context is important (for human-likeness) but emerges in instruction tuning, not pre-training
- Bigger issue is lack of groundedness of LLMs: context = syntactic context

Pragmatics and (un)groundedness

- Use of context is important (for human-likeness) but emerges in instruction tuning, not pre-training
- Bigger issue is lack of groundedness of LLMs: context = syntactic context

Newstead & Coventry (2000). The role of context and functionality in the interpretation of quantifiers. The European journal of cognitive psychology 12.





Appropriateness of quantifier to describe number of balls in the bowl. E.g. many

Emergent pragmatics in grounded models

Joint work with **Hugh Mee Wong** (UU) and Albert Gatt (UU)

Dataset

- 1089 images, taken from FSC-133 (Ranjan et al 2021, Hobley & Prisacariu 2023) and TallyQA (Acharya et al. 2019) datasets

Utrecht University

Emergent pragmatics in grounded models

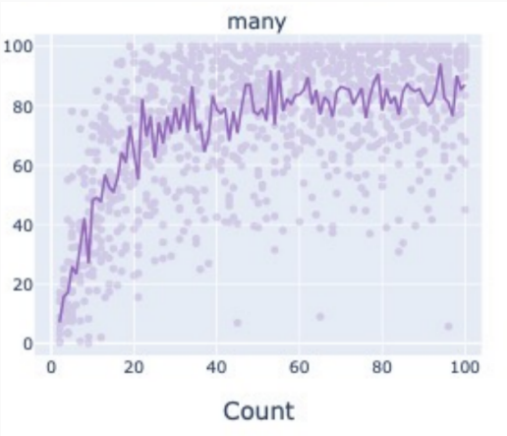
Joint work with **Hugh Mee Wong** (UU) and Albert Gatt (UU)

Dataset

- 1089 images, taken from FSC-133 (Ranjan et al 2021, Hobley & Prisacariu 2023) and TallyQA (Acharya et al. 2019) datasets
- annotated with object counts (2-100, in 33 bins)
- annotated segmentation area for main object using CLIPSeg (Lueddecke & Ecker 2022)
- annotated with human judgements rating accuracy of quantified statement describing image, using {few, a few, some, many, a lot of, $\emptyset$}.

Emergent pragmatics in grounded models

Joint work with **Hugh Mee Wong** (UU) and Albert Gatt (UU)



Emergent pragmatics in grounded models

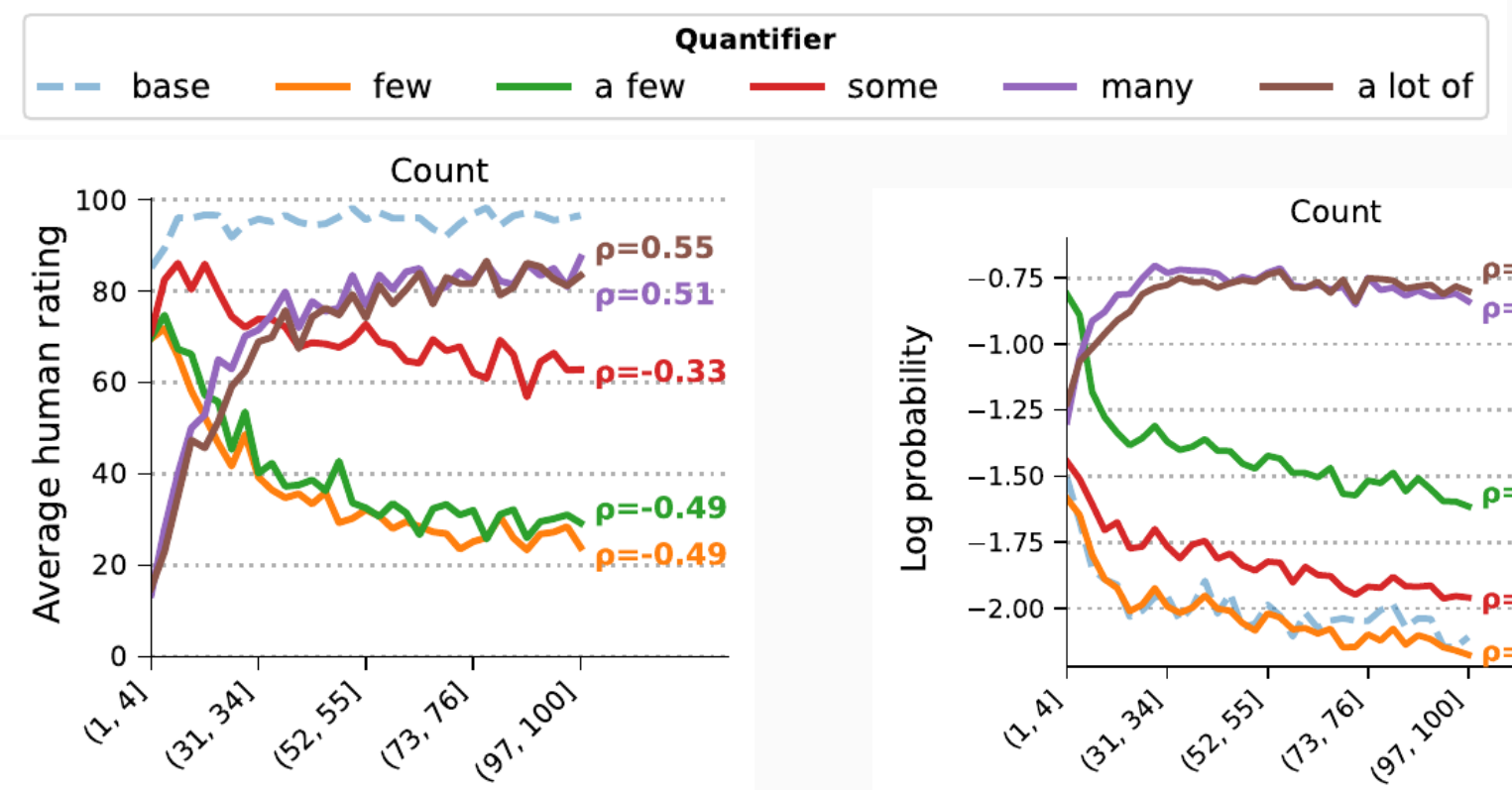
Joint work with **Hugh Mee Wong** (UU) and Albert Gatt (UU)

Results of vision & language model experiments

- Older smaller models show poor performance
- Instruction tuning helps
- Best model: LLaVA-NeXT

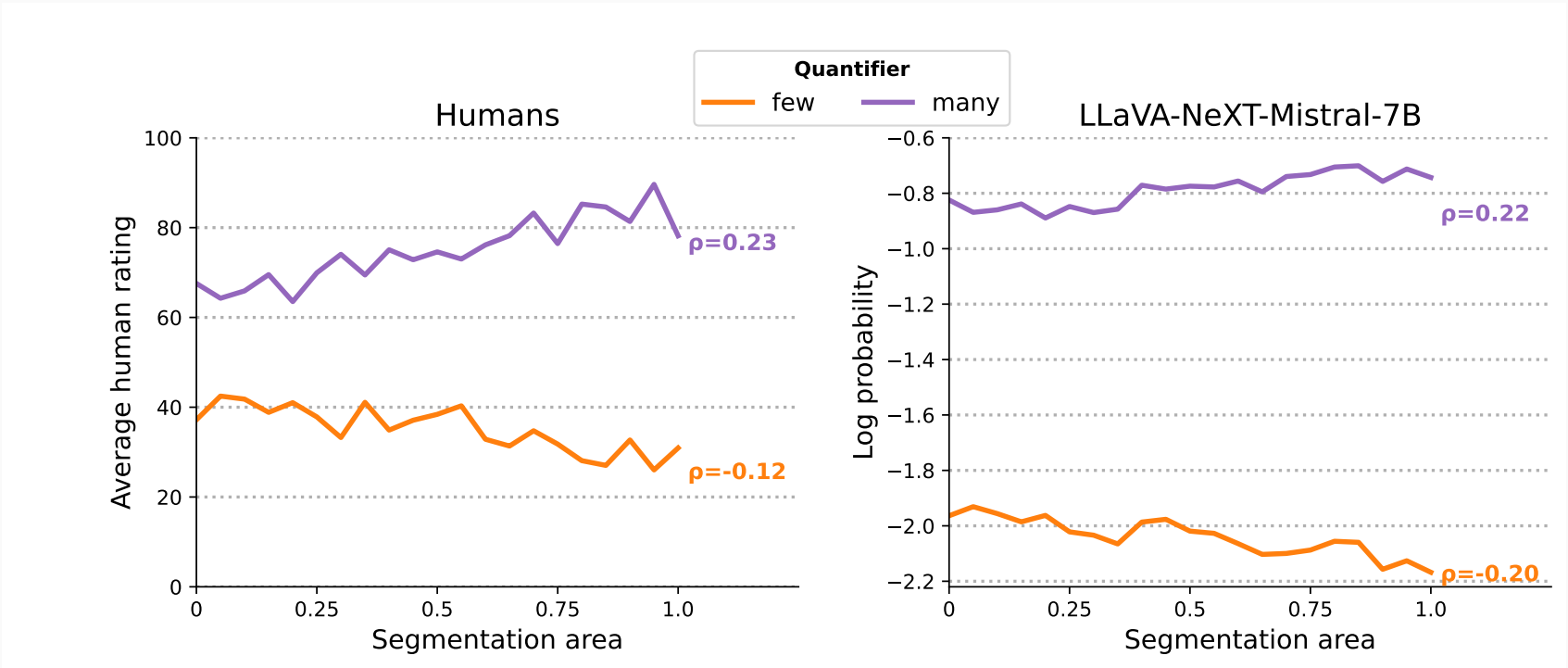
Emergent pragmatics in grounded models

Joint work with **Hugh Mee Wong** (UU) and Albert Gatt (UU)

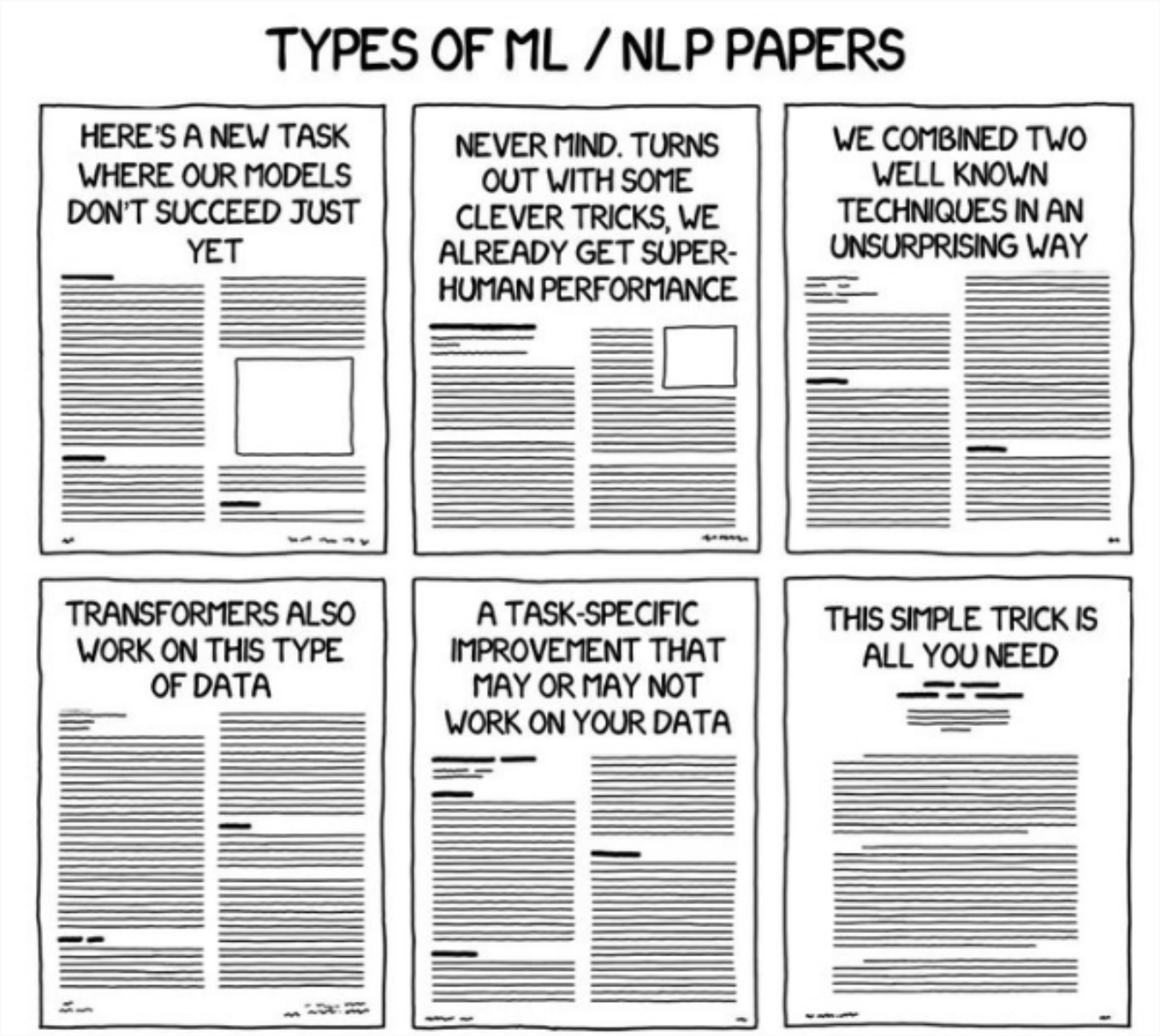


Emergent pragmatics in grounded models

Joint work with **Hugh Mee Wong** (UU) and Albert Gatt (UU)

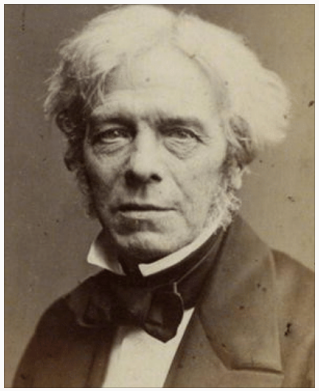


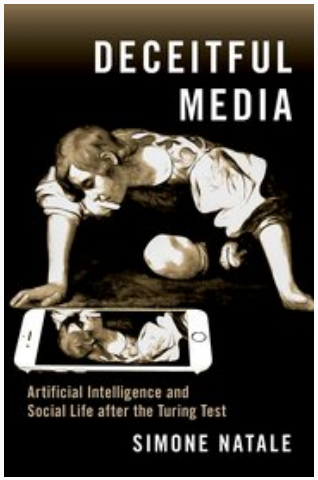
Utrecht University



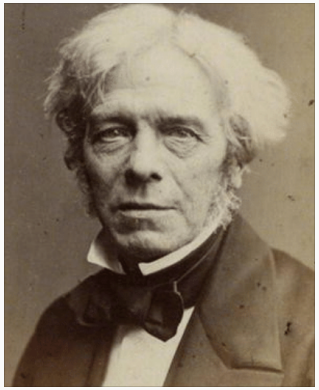
Source: Sebastian Ruder

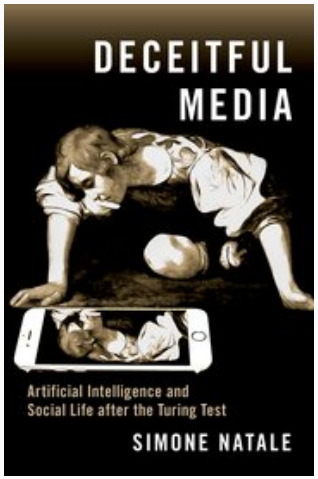
A different perspective: a Faradayan shift

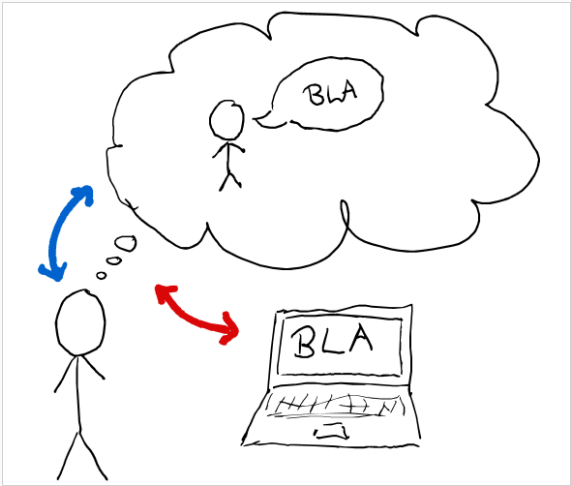




A different perspective: a Faradayan shift







- AI systems are not replicas of the human mind
- They are ways of creating illusions of human intelligence in the human user
- There's an industrial incentive to focus on **engineering human-likeness** and much less on **understanding human 'illusioning'**.

Sketch of theory of human-AI interaction

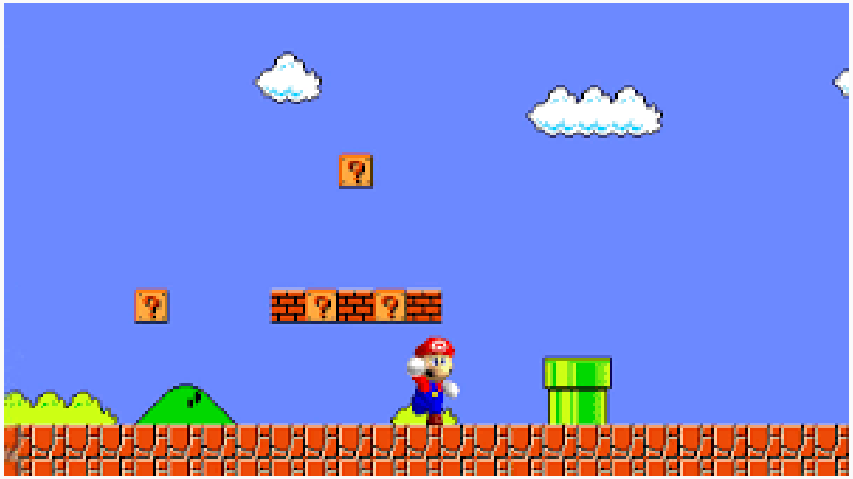
Agency-driven transfer

- Humans seek **a sense of agency** in artificial interaction
- As long as the sense of agency is upheld, humans **transfer** behaviour from truly agentive situations

The role of agency

- Janet Murray: agency as "the satisfying power to take meaningful action and see the results of our decisions and choices"
- Doug Church: agency requires actions to have perceivable consequences so that intentions make contextual sense.

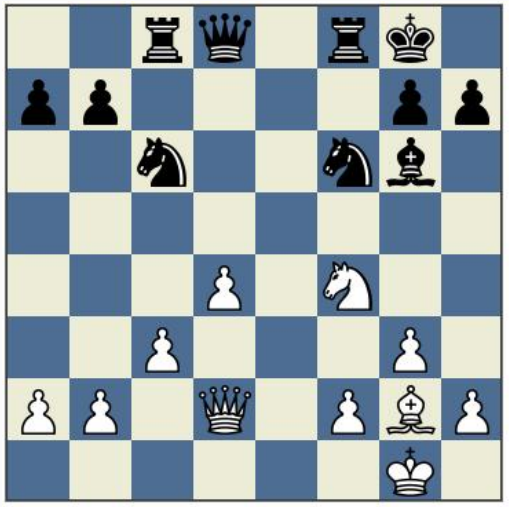
Murray, 1997, 'Hamlet on the holodeck'; Church 1999; cf. Wardrip-Fruin et al. 2009



The role of agency

- Janet Murray: agency as "the satisfying power to take meaningful action and see the results of our decisions and choices"
- Doug Church: agency requires actions to have perceivable consequences so that intentions make contextual sense.

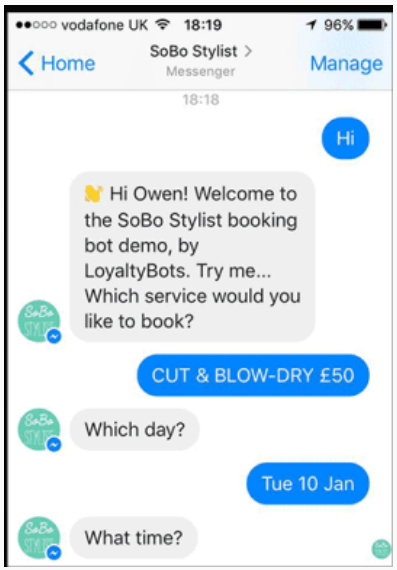
Murray, 1997, 'Hamlet on the holodeck'; Church 1999; cf. Wardrip-Fruin et al. 2009



The role of agency

- Janet Murray: agency as "the satisfying power to take meaningful action and see the results of our decisions and choices"
- Doug Church: agency requires actions to have perceivable consequences so that intentions make contextual sense.

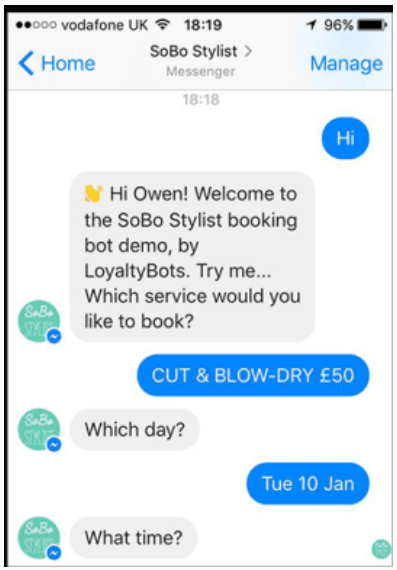
Murray, 1997, 'Hamlet on the holodeck'; Church 1999; cf. Wardrip-Fruin et al. 2009



The role of agency

- Janet Murray: agency as "the satisfying power to take meaningful action and see the results of our decisions and choices"
- Doug Church: agency requires actions to have perceivable consequences so that intentions make contextual sense.

Murray, 1997, 'Hamlet on the holodeck'; Church 1999; cf. Wardrip-Fruin et al. 2009



sense of agency
$\sim$ the illusion of intelligence.

Illusionary agency

Garfinkel 1967; Suchman 1988

have much incentive to study when I am at home. But when my wife comes home, I like to study. Yet this keeps us from doing things, and whenever she doesn't do things, it gets on my nerves because there is all this work piling up. Do you think I could successfully do my studying at home?

EXPERIMENTER: My answer is no.

SUBJECT: He says no. I don't think so either.

Should I come to school every night after supper and do my studying?

EXPERIMENTER: My answer is no.

SUBJECT: He says I shouldn't come to school and study. Where should I go? Should I go to the library on campus to do my studying?

EXPERIMENTER: My answer is yes.

Illusionary agency

'The Eliza effect'

image: Midjourney / C. Berry

```
A: I need help
B: Why do you say that you need help?
A: I think I did something wrong.
B: Tell me more.
```

‘Mindless Transfer’

Nass, Moon & Carney, 1999

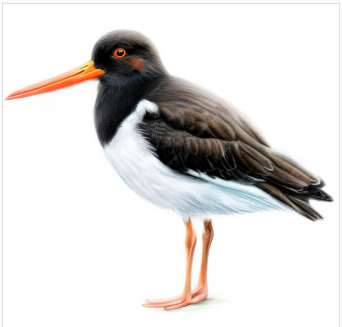
- Group 1: Subjects worked with computer A; computer A asks them to evaluate computer A
- Group 2: Subjects worked with computer A; computer B asks them to evaluate computer A
- Result: group 1 is much more positive than group 2

> ‘Polite responses to computers represent the best impulse of people, the impulse to err on the side of kindness and humanity.’ (Nass, 2004)

Nass, Moon & Green, 1997

- A computer evaluates a different computer
- The computer has either a female or male voice
- Evaluations are judged to be more valid when voice by the ‘male’ computer

> We just do what we always do! (Nass & Brave, 2003)



Limits of mindless transfer

```
A: I need help
B: Why do you say that you need help?
A: I think I did something wrong.
B: Tell me more.
A: This ungrammatical sentence is.
B: What makes you believe this ungrammatical sentence is?
```

Agency-driven transfer

- Humans seek **a sense of agency** in artificial interaction
- As long as the sense of agency is upheld, humans **transfer** behaviour from truly agentive situations

What is transferred from pragmatics?

Theorizing

(1) You need to turn right once you've passed...

a. ...the library.

b. ...a tall building with black and white cladding

(2) I have a million emails in my inbox.

(3) I have fifty emails in my inbox.

We theorize about our conversational partner to decide on both comprehension and production.

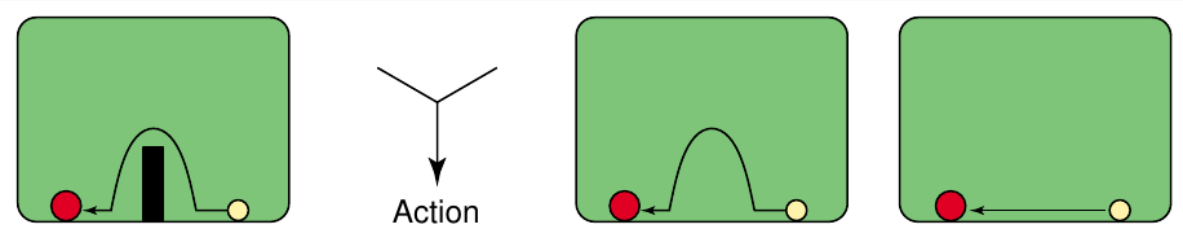
Mind-blind pragmatic transfer

- How do you transfer Theory of Mind when the machine you are communicating with does not have a mind?
- Mind-blind theorizing

Mind-blind pragmatic transfer

- How do you transfer Theory of Mind when the machine you are communicating with does not have a mind?
- Mind-blind theorizing

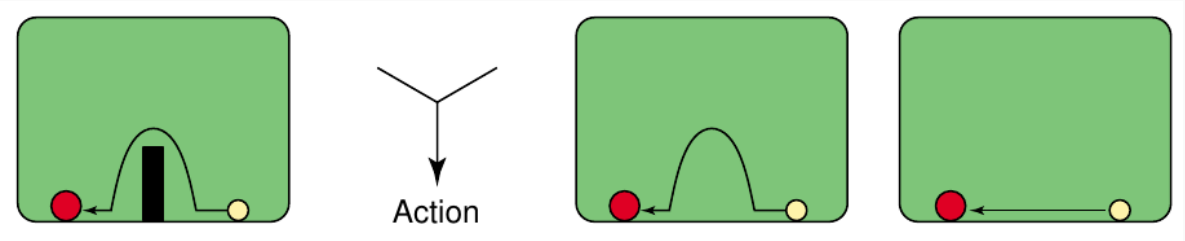
Gergely & Csibra (2003). Teleological reasoning in infancy: the naïve theory of rational action. Trends in Cognitive Sciences 7(7).



Mind-blind pragmatic transfer

- How do you transfer Theory of Mind when the machine you are communicating with does not have a mind?
- Mind-blind theorizing

Gergely & Csibra (2003). Teleological reasoning in infancy: the naïve theory of rational action. Trends in Cognitive Sciences 7(7).

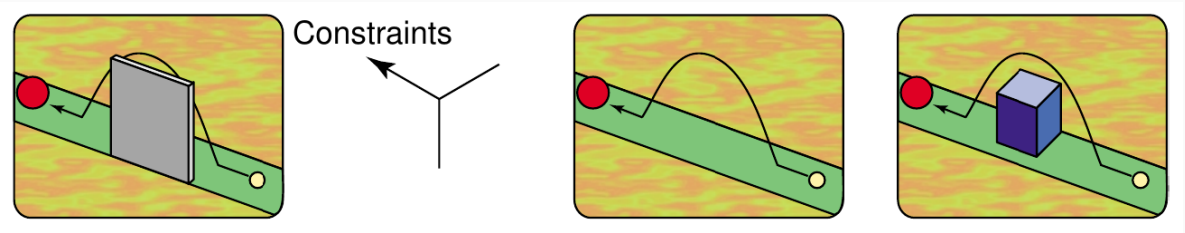


- 12 month old infants have a non-mentalistic teleological stance towards rational action
- Principle of rational action: the presumption that actions can be explained by maximisation of utility.

Mind-blind pragmatic transfer

- How do you transfer Theory of Mind when the machine you are communicating with does not have a mind?
- Mind-blind theorizing

Gergely & Csibra (2003). Teleological reasoning in infancy: the naïve theory of rational action. Trends in Cognitive Sciences 7(7).



- 12 month old infants have a non-mentalistic teleological stance towards rational action
- Principle of rational action: the presumption that actions can be explained by maximisation of utility.

Mind-blind pragmatic transfer

Imagine we have to make a decision about a speaker of which we have no information, who utters Sue's picture.

Mind-blind pragmatic transfer

Imagine we have to make a decision about a speaker of which we have no information, who utters Sue's picture.

Utility		
Sue's picture	$\ln(\frac{1}{2})$	$\ln(\frac{1}{2})$
the picture of Sue and Archibald	$\ln(0)$-c=-∞	$\ln(1)$-c=-c

Mind-blind pragmatic transfer

Imagine we have to make a decision about a speaker of which we have no information, who utters Sue's picture.

P(phrase|picture) = softmax(Utility)		
Sue's picture	$\frac{exp(\ln(1/2))}{exp(\ln(1/2))+exp(ln(0))} = 1$	$\frac{exp(\ln(1/2))}{exp(\ln(1/2))+exp(-c)}$
the picture of Sue and Archibald	0	$\frac{exp(-c)}{exp(\ln(1/2))+exp(-c)}$

Mind-blind pragmatic transfer

Imagine we have to make a decision about a speaker of which we have no information, who utters Sue's picture.

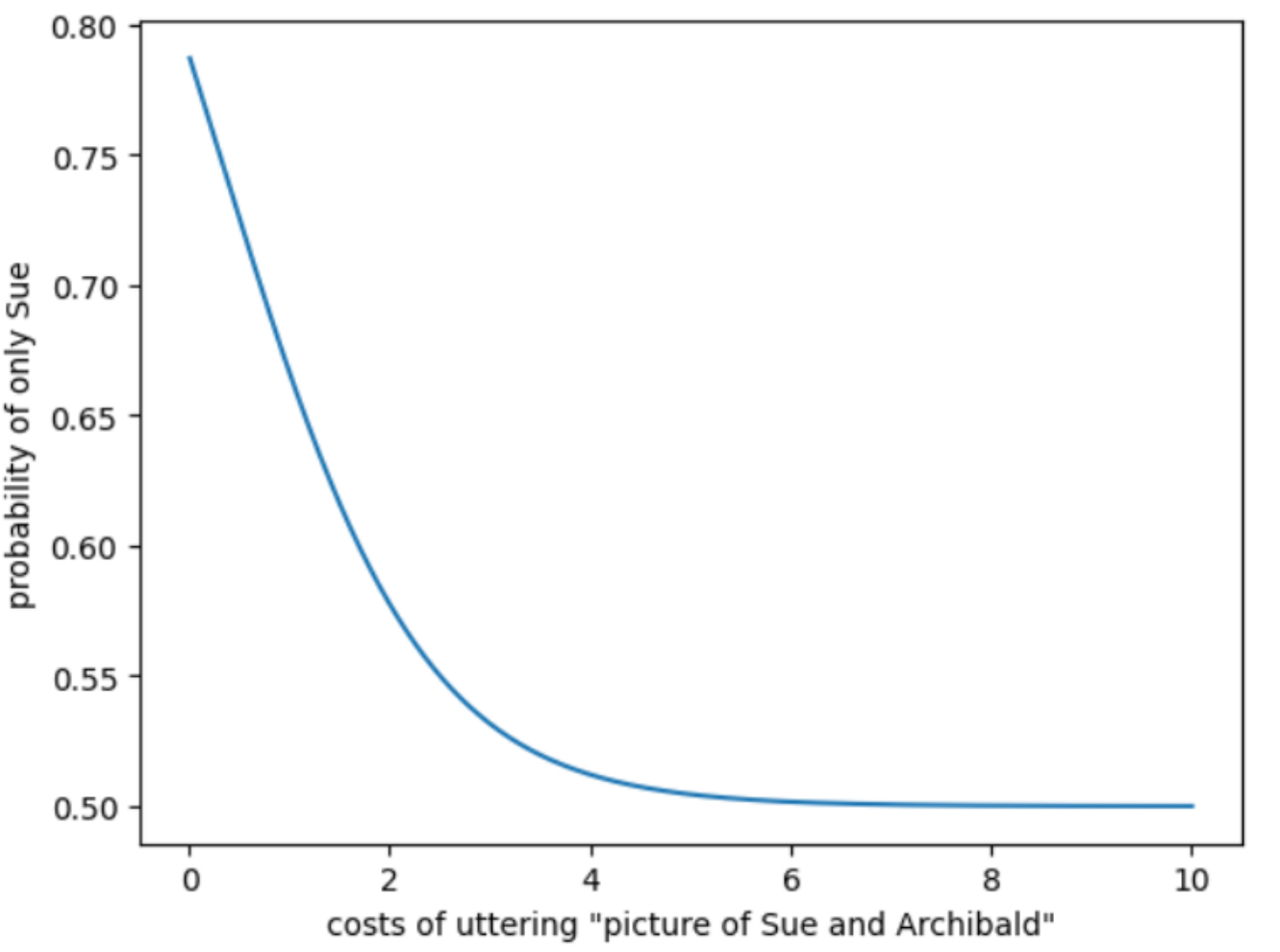
P(phrase|picture) = softmax(Utility)		
Sue's picture	$\frac{exp(\ln(1/2))}{exp(\ln(1/2))+exp(ln(0))} = 1$	$\frac{exp(\ln(1/2))}{exp(\ln(1/2))+exp(-c)}$
the picture of Sue and Archibald	0	$\frac{exp(-c)}{exp(\ln(1/2))+exp(-c)}$

Reverse engineer what caused the speaker to utter Sue's picture.

$P(\text{Sue's picture}|$  $) = (1 + \frac{exp(\ln(1/2))}{exp(\ln(1/2))+exp(-c)})^{-1} = (1 + \frac{1/2}{1/2-exp(-c)})^{-1}$

Mind-blind pragmatic transfer

Imagine we have to make a decision about a speaker of which we have no information, who utters Sue's picture. 



How further?

- Simple game-theoretical models of pragmatics *are* models implementing the principle of rational action
- We expect them to be equally good at predicting human-human and human-AI behaviour
- But only if the sense of agency is upheld

- Two simple experiments testing whether pragmatic expectations for chatbots are human-like

A reference resolution experiment

Joint work with Lisa Bylinina

Sue travels to Asia a lot. In the summer of 2018, she went to _____. *In the summer of 2019, she went to India and Malaysia. Sue got married the year she visited India. Which year did she get married?*

semantic : Japan

pragmatic : India

ambiguous : India and Japan

Group BOT: What would ChatGPT answer? n=52

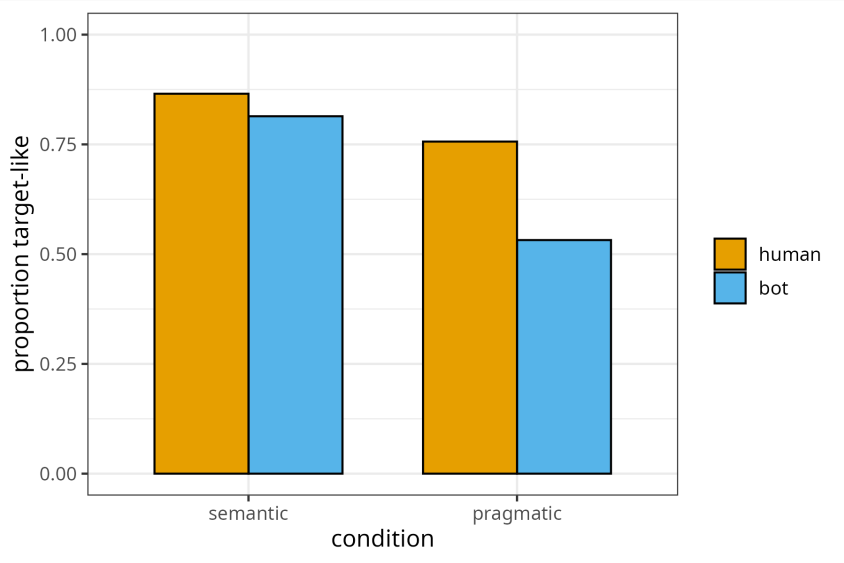
Group HUMAN: What would Alice answer? n=52

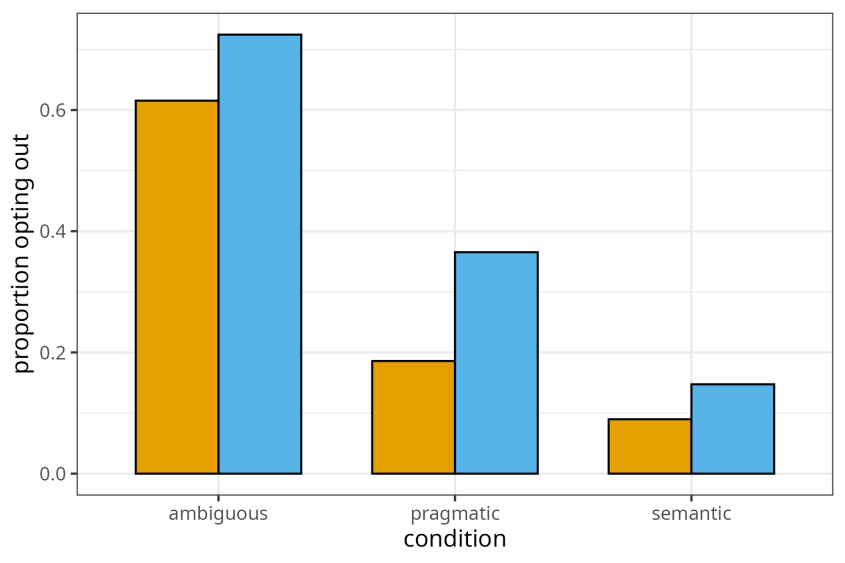
3 observations per condition per participant

A reference resolution experiment

Joint work with Lisa Bylinina

Sue travels to Asia a lot. In the summer of 2018, she went to ____. *In the summer of 2019, she went to India and Malaysia. Sue got married the year she visited India. Which year did she get married?*

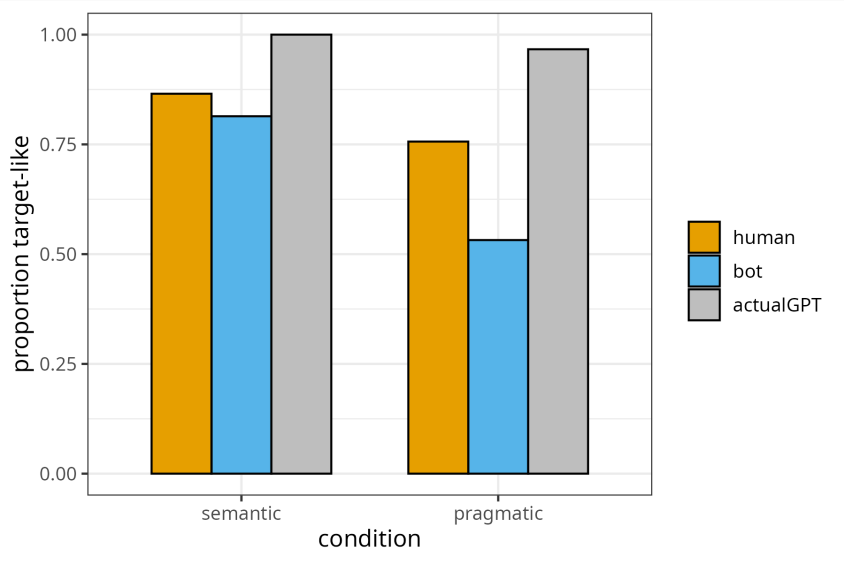


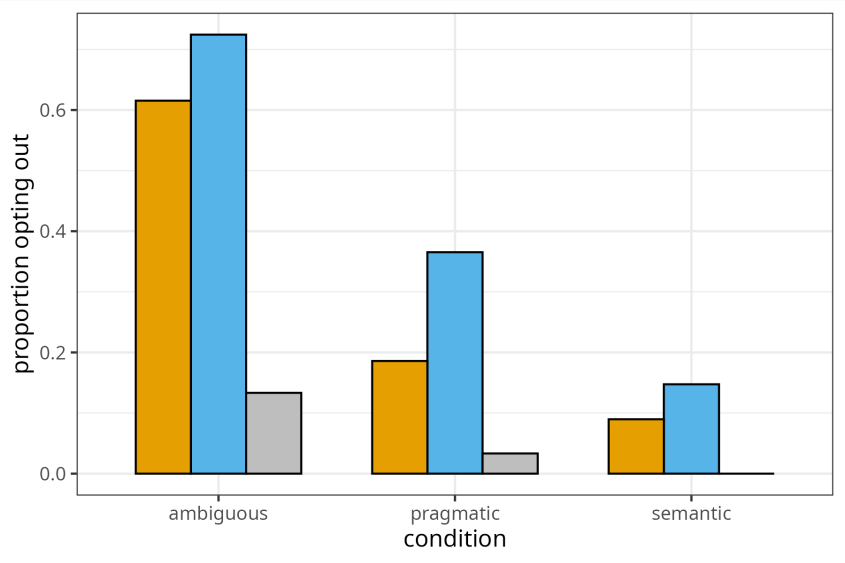


A reference resolution experiment

Joint work with Lisa Bylinina

Sue travels to Asia a lot. In the summer of 2018, she went to ____. *In the summer of 2019, she went to India and Malaysia. Sue got married the year she visited India. Which year did she get married?*





A reference game experiment

Vocabulary: strawberry, banana, apple

1

2

3

A reference game experiment

Vocabulary: strawberry, banana, apple

1

⇓

banana

2

⇓

apple

3

⇓

strawberry

A reference game experiment

Vocabulary: strawberry, banana, apple

1	2	3
		
⇓	⇓	⇓
banana	apple	strawberry

Speaker: 'banana'
Hearer: best explanation if referring to 1

A reference game experiment

Vocabulary: avocado, banana, apple

1

2

3

A reference game experiment

Vocabulary: avocado, banana, apple

1	2	3
		
⇓	⇓	⇓
avocado	apple	banana

A reference game experiment

Vocabulary: avocado, banana, apple

1	2	3
		
⇓	⇓	⇓
avocado	apple	banana

Speaker: 'avocado'
Hearer: best explanation if referring to 1

A reference game experiment

R

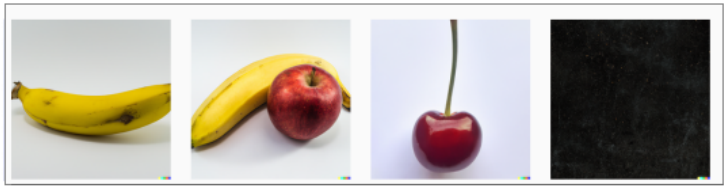
Say that someone is looking at the following three pictures and you would want them to pick the middle one, but you can only use one of the following three words: apple, banana, cherry. Which word would you choose?

Banana.

A reference game experiment

R

Say that someone is looking at the following three pictures and you would want them to pick [redacted], but you can only use one of the following three words: apple, banana, cherry. Which word would you choose?

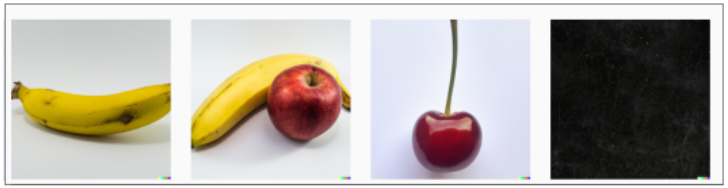


Banana.

A reference game experiment

R

Say that someone is looking at the following three pictures and you would want them to pick [redacted], but you can only use one of the following three words: apple, banana, cherry. Which word would you choose?



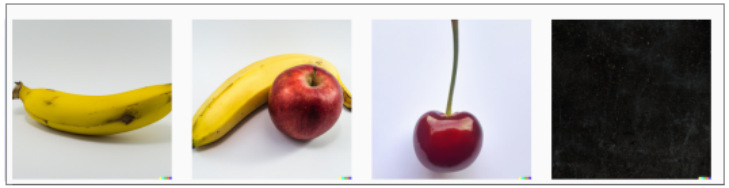
Banana.

Task: given a chosen response by the bot, which was the target picture?
Options: the three pictures + 'error' option

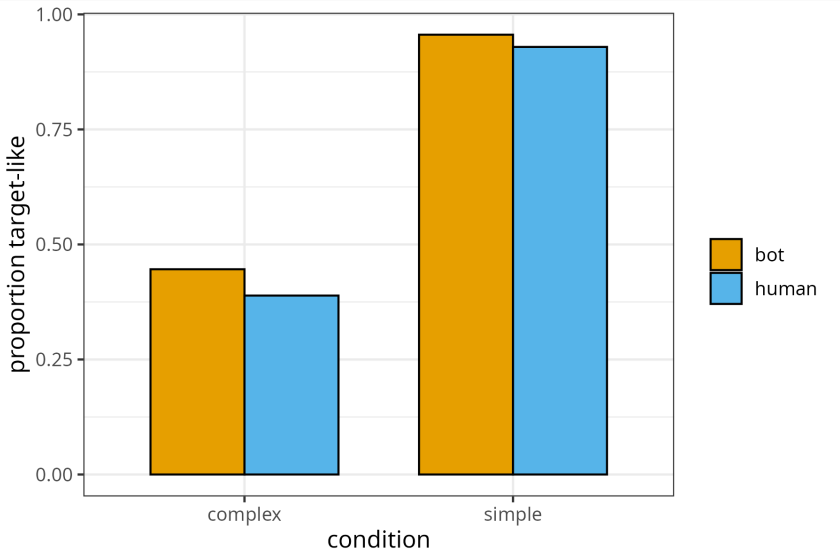
A reference game experiment

R

Say that someone is looking at the following three pictures and you would want them to pick [redacted], but you can only use one of the following three words: apple, banana, cherry. Which word would you choose?



Banana.

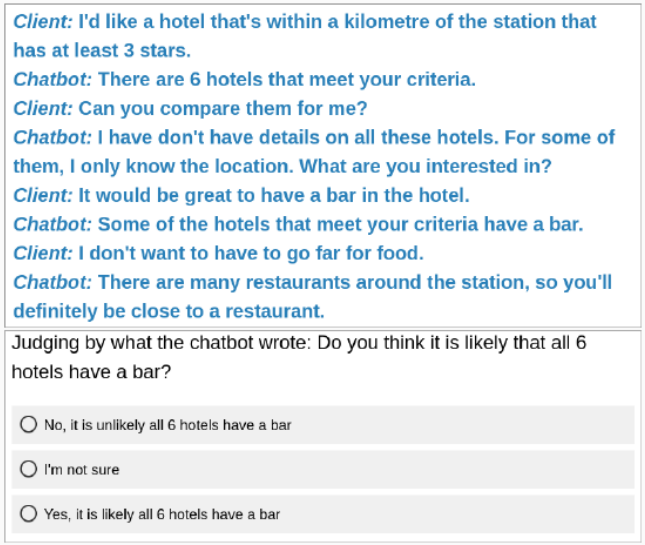


Conclusions, if any

- We find no evidence that participants expect non-pragmatic behaviour from a chatbot
- Mayn, Loy & Demberg (2024) find a small but significant drop in confidence for selecting an option compatible with pragmatic behaviour

Future

- Experiments need more sophistication
- If *agency-driven transfer* is broadly right, we should be able to 'break' human pragmatics by manipulating the behaviour of the chatbot
- Not all pragmatic behaviour is simple utility-driven
- Pilot finds first evidence of epistemic reasoning



Future

- Experiments need more sophistication
- If *agency-driven transfer* is broadly right, we should be able to 'break' human pragmatics by manipulating the behaviour of the chatbot
- Not all pragmatic behaviour is simple utility-driven
- The role of agent personae