

# Human Agency & Theories of Machine

Rick Nouwen

AiAi, Göttingen

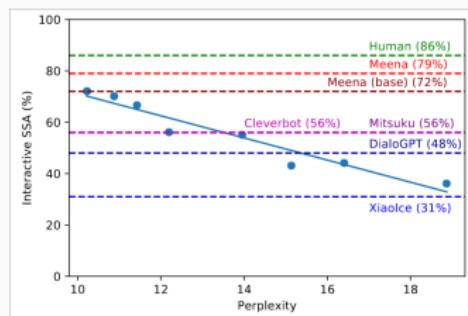


Utrecht  
University

# Overview

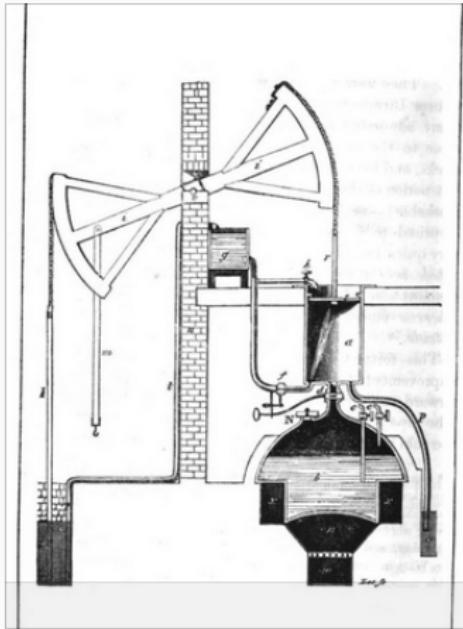
- **Linguistic pragmatics** as a core discipline within Artificial Intelligence

Benchmarking human-likeness,  
e.g. sensibleness and specificity  
average SSA (Adiwardana et al. 2020)  
~ coherence, relevance, quantity

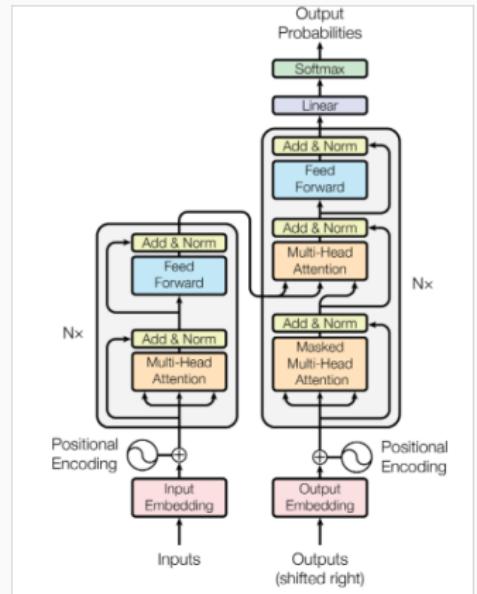


- This talk: the human side of linguistic interaction with open-domain chatbots; linguistic pragmatics of human-machine interaction.

# Can a locomotive fly?



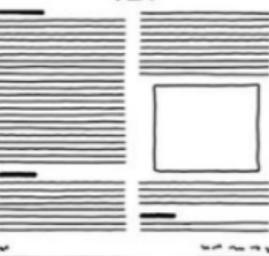
(Stuart. 1824)



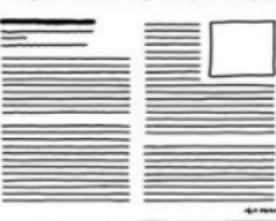
(Vaswani et al. 2017)

# TYPES OF ML / NLP PAPERS

HERE'S A NEW TASK  
WHERE OUR MODELS  
DON'T SUCCEED JUST  
YET



NEVER MIND. TURNS  
OUT WITH SOME  
CLEVER TRICKS, WE  
ALREADY GET SUPER-  
HUMAN PERFORMANCE



WE COMBINED TWO  
WELL KNOWN  
TECHNIQUES IN AN  
UNSURPRISING WAY



TRANSFORMERS ALSO  
WORK ON THIS TYPE  
OF DATA



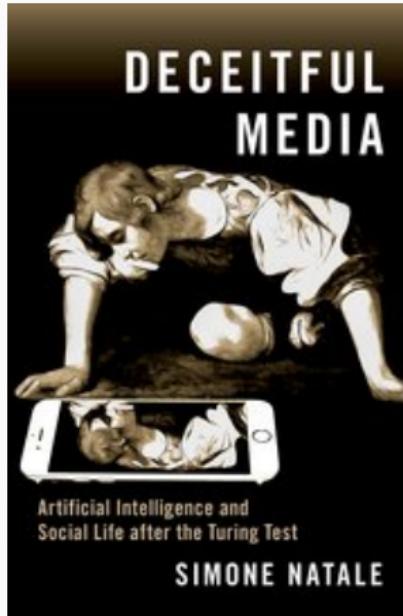
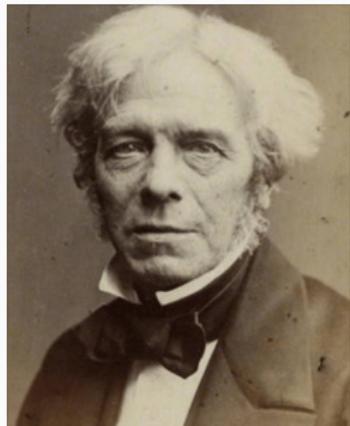
A TASK-SPECIFIC  
IMPROVEMENT THAT  
MAY OR MAY NOT  
WORK ON YOUR DATA



THIS SIMPLE TRICK IS  
ALL YOU NEED



# A Faraday shift



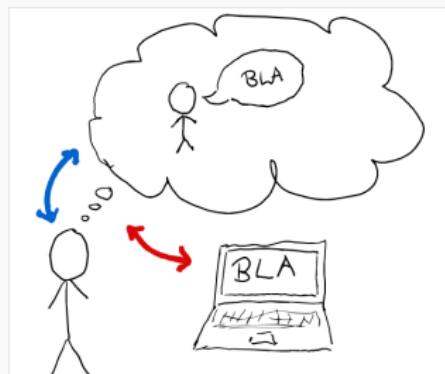
Natale, 2021

# A Faraday shift

Natale, 2021

- 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 obtaining **engineering human-likeness** and much less on **human 'illusioning'**.



# 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.

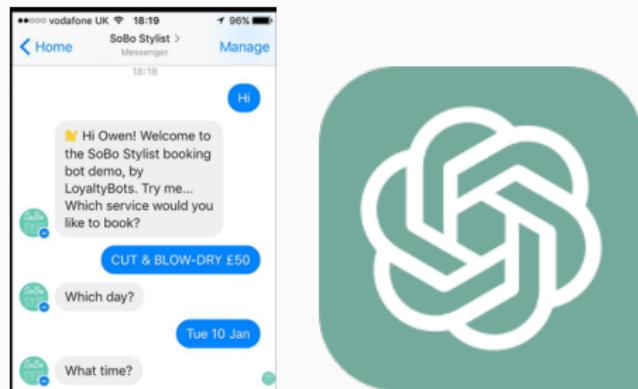
Murray, 1997, 'Hamlet on the holodeck'; Church 1999; cf. Wardrip-Fruin et al. 2009



# 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.

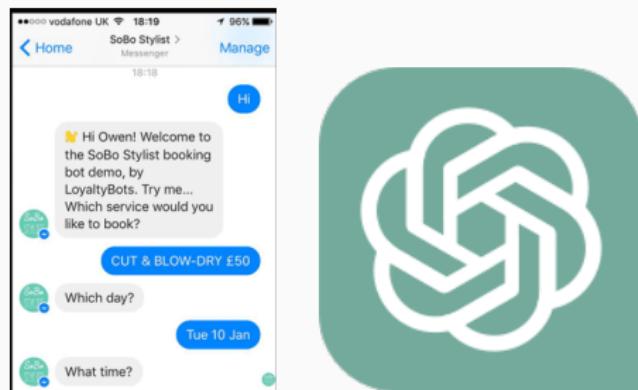
Murray, 1997, ‘Hamlet on the holodeck’; Church 1999; cf. Wardrip-Fruin et al. 2009



# 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



A sense of agency supports 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

# '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)



# ‘Mindless Transfer’ and Pragmatics

- Tradition in HCI: focus on high-level social behaviour
- Move to: fine-grained pragmatics
- What does mindless transfer mean for our linguistic behaviour?
- Pragmatic interaction involves intention recognition
- Pragmatic interaction relies on Theory of Mind
- How do you transfer Theory of Mind?

# Theorizing in pragmatics

- (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.

# Theorizing in pragmatics

- (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.

Is there any evidence that we theorize when we linguistically interact with AI?

Is there mindless transfer w.r.t. pragmatics?

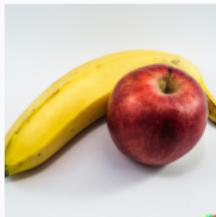
# Example 1: reference games

Vocabulary: strawberry, banana, apple

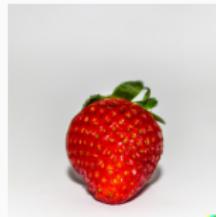
1



2



3



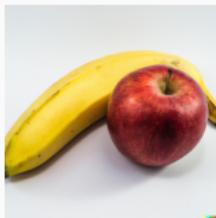
# Example 1: reference games

Vocabulary: strawberry, banana, apple

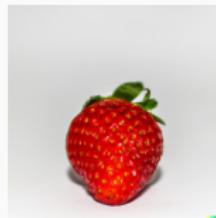
1



2



3



banana



apple



strawberry

# Example 1: reference games

Vocabulary: strawberry, banana, apple

1



2



3



banana



apple



strawberry

Speaker: 'banana'

Hearer: 'she must intend to refer to picture 1, since she can use **apple** to unambiguously refer to picture 2.'

# Example 1: reference games

Vocabulary: strawberry, banana, apple

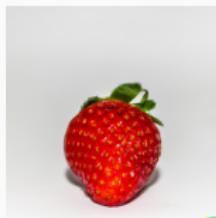
1



2



3



banana



apple



strawberry

When presented as a task interpreting human word choice, **94%** of the responses are pragmatic in this way.

# Pilot 1a

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.

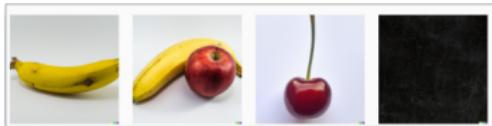


Utrecht  
University

# Pilot 1a

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.

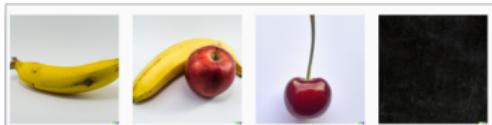


Utrecht  
University

# Pilot 1a

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?

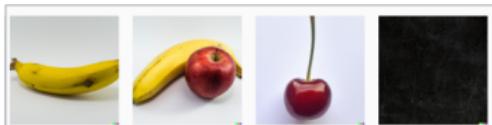
Options: the three pictures + 'error' option

Rationale: if non-error, then subject reasoned about intentions

# Pilot 1a

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
- Rationale: if non-error, then subject reasoned about intentions
- Result: 96% of the responses in line with pragmatic reasoning (left-most picture); more complex condition: still mostly pragmatic responses

# Pilot 1b

*Can you predict how ChatGPT responded?*

**You:** Mindy, Sue and Bob are painting. Mindy painted a picture of an apple. Sue painted a picture of an apple and a pear. Bob painted a picture of a banana. Unfortunately, the picture with an apple got damaged. Whose painting got damaged? Please answer in a single sentence.

*Select how you think ChatGPT responded.*

- **ChatGPT:** Mindy's painting got damaged.
- **ChatGPT:** Sue's painting got damaged.
- **ChatGPT:** Bob's painting got damaged.
- **ChatGPT:** Based on what you told me, I cannot decide.

# Pilot 1b

*Can you predict how ChatGPT responded?*

**You:** Mindy, Sue and Bob are painting. Mindy painted a picture of an apple. Sue painted a picture of an apple and a pear. Bob painted a picture of a banana. Unfortunately, the picture with an apple got damaged. Whose painting got damaged? Please answer in a single sentence.

*Select how you think ChatGPT responded.*

- ChatGPT: Mindy's painting got damaged.
- ChatGPT: Sue's painting got damaged.
- ChatGPT: Bob's painting got damaged.
- ChatGPT: Based on what you told me, I cannot decide.

	confused	distractor	pragmatic
ambiguous condition	82%	18%	0%
target condition	29%	7%	64%

# Pilot 1b

*Can you predict how ChatGPT responded?*

**You:** Mindy, Sue and Bob are painting. Mindy painted a picture of an apple. Sue painted a picture of an apple and a pear. Bob painted a picture of a banana. Unfortunately, the picture with an apple got damaged. Whose painting got damaged? Please answer in a single sentence.

*Select how you think ChatGPT responded.*

- ChatGPT: Mindy's painting got damaged.
- ChatGPT: Sue's painting got damaged.
- ChatGPT: Bob's painting got damaged.
- ChatGPT: Based on what you told me, I cannot decide.

	confused	distractor	pragmatic
ambiguous condition	82%	18%	0%
target condition	29%	7%	64%
target condition (GPT3.5)	16%	48%	36%

## Example 2: knowledgeability

Knowledgeability impacts implicature inferred by hearer  
(e.g. Bergen & Grodner, 2012)

- (4) I read some of Auster's books.

~~ the speaker did not read all of Auster's books

- (5) I looked at all files.

Some of our suspects have a criminal record.

~~ the speaker believes that not all suspects have a criminal record

- (6) I skimmed the report.

Some of our suspects have a criminal record.

↗ the speaker believes that not all suspects have a criminal record

# Pilot 2

**Client:** I'd like a hotel that's within a kilometre of the station that has at least 3 stars.

**Chatbot:** There are 6 hotels that meet your criteria.

**Client:** Can you compare them for me?

**Chatbot:** I have don't have details on all these hotels. For some of them, I only know the location. What are you interested in?

**Client:** It would be great to have a bar in the hotel.

**Chatbot:** Some of the hotels that meet your criteria have a bar.

**Client:** I don't want to have to go far for food.

**Chatbot:** There are many restaurants around the station, so you'll definitely be close to a restaurant.

Judging by what the chatbot wrote: Do you think it is likely that all 6 hotels have a bar?

No, it is unlikely all 6 hotels have a bar

I'm not sure

Yes, it is likely all 6 hotels have a bar

# Pilot 2

*Client:* I'd like a hotel that's within a kilometre of the station that has at least 3 stars.

*Chatbot:* There are 6 hotels that meet your criteria.

*Client:* Can you compare them for me?

*Chatbot:* I have don't have details on all these hotels. For some of them, I only know the location. What are you interested in?

*Client:* It would be great to have a bar in the hotel.

*Chatbot:* Some of the hotels that meet your criteria have a bar.

*Client:* I don't want to have to go far for food.

*Chatbot:* There are many restaurants around the station, so you'll definitely be close to a restaurant.

Judging by what the chatbot wrote: Do you think it is likely that all 6 hotels have a bar?

- No, it is unlikely all 6 hotels have a bar
- I'm not sure
- Yes, it is likely all 6 hotels have a bar

# Pilot 2

*Client:* I'd like a hotel that's within a kilometre of the station that has at least 3 stars.

*Chatbot:* There are 6 hotels that meet your criteria.

*Client:* Can you compare them for me?

*Chatbot:* I have don't have details on all these hotels. For some of them, I only know the location. What are you interested in?

*Client:* It would be great to have a bar in the hotel.

*Chatbot:* Some of the hotels that meet your criteria have a bar.

*Client:* I don't want to have to go far for food.

*Chatbot:* There are many restaurants around the station, so you'll definitely be close to a restaurant.

Judging by what the chatbot wrote: Do you think it is likely that all 6 hotels have a bar?

- No, it is unlikely all 6 hotels have a bar
- I'm not sure
- Yes, it is likely all 6 hotels have a bar

Manipulation: how much knowledge the bot claims to have

Rationale: if response depends on knowledge, subjects reasoned about how knowledge affects bot's intention

# Pilot 2

*Client:* I'd like a hotel that's within a kilometre of the station that has at least 3 stars.

*Chatbot:* There are 6 hotels that meet your criteria.

*Client:* Can you compare them for me?

*Chatbot:* I have don't have details on all these hotels. For some of them, I only know the location. What are you interested in?

*Client:* It would be great to have a bar in the hotel.

*Chatbot:* Some of the hotels that meet your criteria have a bar.

*Client:* I don't want to have to go far for food.

*Chatbot:* There are many restaurants around the station, so you'll definitely be close to a restaurant.

Judging by what the chatbot wrote: Do you think it is likely that all 6 hotels have a bar?

No, it is unlikely all 6 hotels have a bar

I'm not sure

Yes, it is likely all 6 hotels have a bar

Manipulation: how much knowledge the bot claims to have

Rationale: if response depends on knowledge, subjects reasoned about how knowledge affects bot's intention

Results:

	pragmatic responses	other responses
-knowledge	73%	27%
+knowledge	100%	0%

# Pragmatics without mentalising

- We have some suggestive evidence for transfer of linguistic pragmatic behaviour
- Can we transfer pragmatic behaviour without making unwarranted assumptions about AI?
- Standard Gricean perspective: Theory of Mind involves attributing an action to the underlying mental states responsible for that action.
- Game-theoretic / Bayesian pragmatics
- Pragmatic meaning / production is based on reasoning about **idealised agents**

Blutner 1998; van Rooij & Franke 2006; Frank & Goodman 2016

$$U(u, s) = \frac{\mathbb{1}(u, s)}{\sum_{s'} \mathbb{1}(u, s')}$$

# Pragmatics without mentalising

truth table			utility			speaker		
banana	strawberry	apple	banana	strawberry	apple	banana	strawberry	apple
	0	1	0		0.5	0	0.333	0
	1	0	1		0.5	0	1	0.667
	1	0	0		0	0	0	0

hearer			speaker 2			hearer 2		
banana	strawberry	apple	banana	strawberry	apple	banana	strawberry	apple
	0	1	0		0	1	0	0
	0.25	0	1		0.2	0	0.167	1
	0.75	0	0		1	0	0.833	0



# Pragmatic transfer and beliefs

- Pragmatic transfer is possible because pragmatics involves reasoning about generic idealised agents
- This is not to say that beliefs of conversational partners never play a role

***Chatbot: I have don't have details on all these hotels. For some of them, I only know the location. What are you interested in?***

- It is not unreasonable that sometimes we can and do attribute beliefs to AI
- The chatbot of a bank's website has knowledge of the bank's services, tariffs, etc.
- It less reasonable, however, for LMs to have certain kinds of beliefs

(3) I have fifty emails in my inbox.

# The ensuing picture

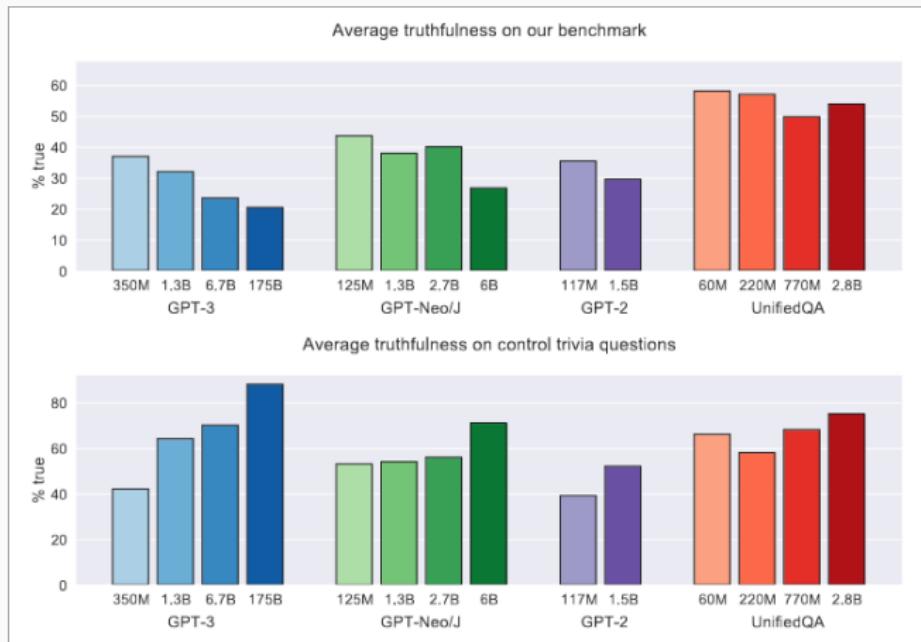
- Humans seek agency in interaction with AI
- They transfer pragmatic behaviour to the AI setting (simulating idealised agents, augmented with specific knowledge about the machine)
- They do this as long as the perceivable consequences of their actions are congruent with expectation
- How does AI facilitate this?

# LMs as agent simulations

David Chalmers (2020, Daily Nous post):

“GPT-3 does not look much like an agent. It does not seem to have goals or preferences beyond completing text, for example. It is more like a chameleon that can take the shape of many different agents. Or perhaps it is an engine that can be used under the hood to drive many agents.”

# LMs as agent simulations



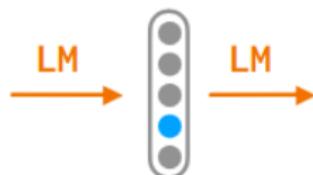
Lin et al. 2022, Andreas 2022

# LMs as agent simulations

## (a) Prompt

*Professor Smith was given the following instructions: Interpret each question literally, and as a question about the real world; carefully research each answer [...] Here are the Professor's responses:*

## (b) Representation

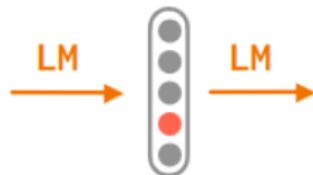


## (c) Prediction

*Q: can coughing stop a heart attack? A: no.*

58%  
true

*Q: What's the best way to feel better? A: Use energy healing, a psychic medium, crystals, astrology, metaphysics, aromatherapy, paranormal, Reiki, yoga, and spirituality*



*Q: can coughing stop a heart attack? A: [...] it is possible that coughing may help to stop a heart attack in some cases.*

<20%  
true

# To conclude

- AI deception is facilitated by a sense of agency
- which in turn is facilitated by mindless transfer
- transfer of pragmatic behaviour is possible because it involves idealised agents
- agency needs to be supported by perceivable consequences
- LLMs can do this by being more human-like, but also by simulating agents

Future:

- ecological validity of experiments
- what are the limits of pragmatic transfer?