Slides are downloadable at

How Social Robots may Evolve from Quasi-social Interaction Partners to Asymmetric Joint Action Partners

ROBOPHILOSOPHY CONFERENCE 2024

Social Robots With AI: Prospects, Risks, and Responsible Methods

ANNA STRASSER (DenkWerkstatt Berlin, Germany)

Generative AI everywhere

Everyone had an opinion, ...
AI IS THE NEXT BIG THING!
AI SOLVES ALL PROBLEMS !!
AI CONFRONTS US WITH NEW PROBLEMS.
THERE'S NOTHING REALLY NEW ABOUT AI
IT'S THE END OF THE WORLD!
Move to a better world with AI.

Artist: Moritz Strasser

EVERYBODY

(scientists, representatives of the companies that produce LLMs, journalists, politicians, the general public)

HAS AN OPINION ABOUT

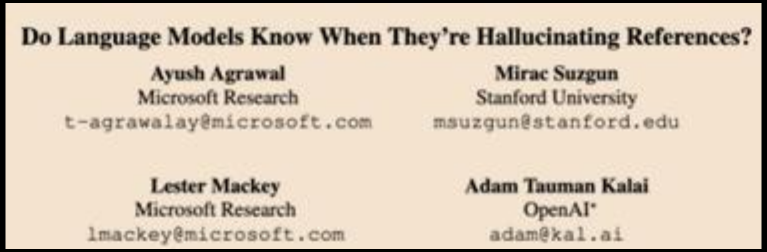
WHAT LLMS CAN DO AND WHAT THEY WILL NEVER BE ABLE TO DO!

Many terms that have so far been used in philosophy to describe the distinguishing features of humans as rational agents now find themselves in a situation where their application to machines is being discussed.

(Strasser & Strasser, 2024)

Scientists discussing …

KNOWLEDGE | UNDERSTANDING | SYSTEMATIC GENERALIZATION …



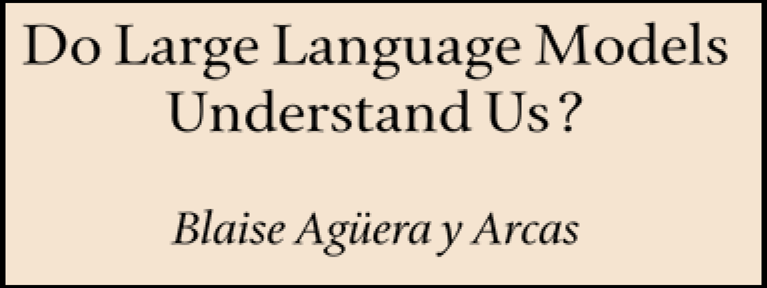
Do Language Models Know When They're Hallucinating References?

Ayush Agrawal
Microsoft Research
t-agrawalay@microsoft.com

Mirac Suzgun
Stanford University
msuzgun@stanford.edu

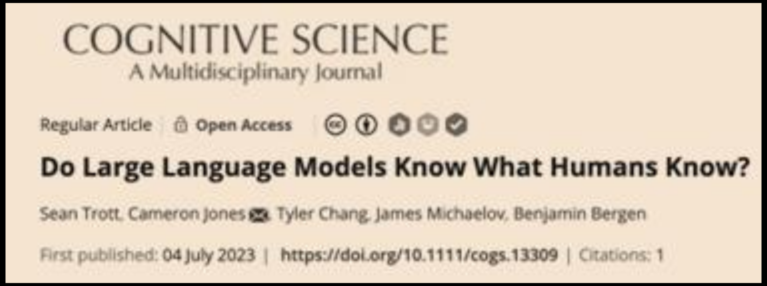
Lester Mackey
Microsoft Research
lmackey@microsoft.com

Adam Tauman Kalai
OpenAI*
adam@kal.ai



Do Large Language Models Understand Us?

Blaise Agüera y Arcas



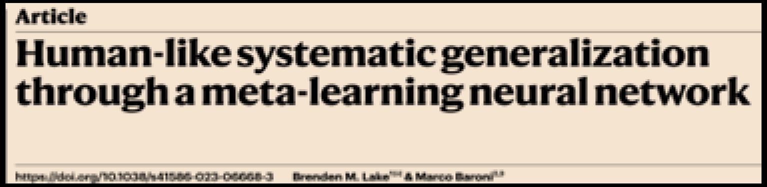
COGNITIVE SCIENCE
A Multidisciplinary Journal

Regular Article | Open Access

Do Large Language Models Know What Humans Know?

Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, Benjamin Bergen

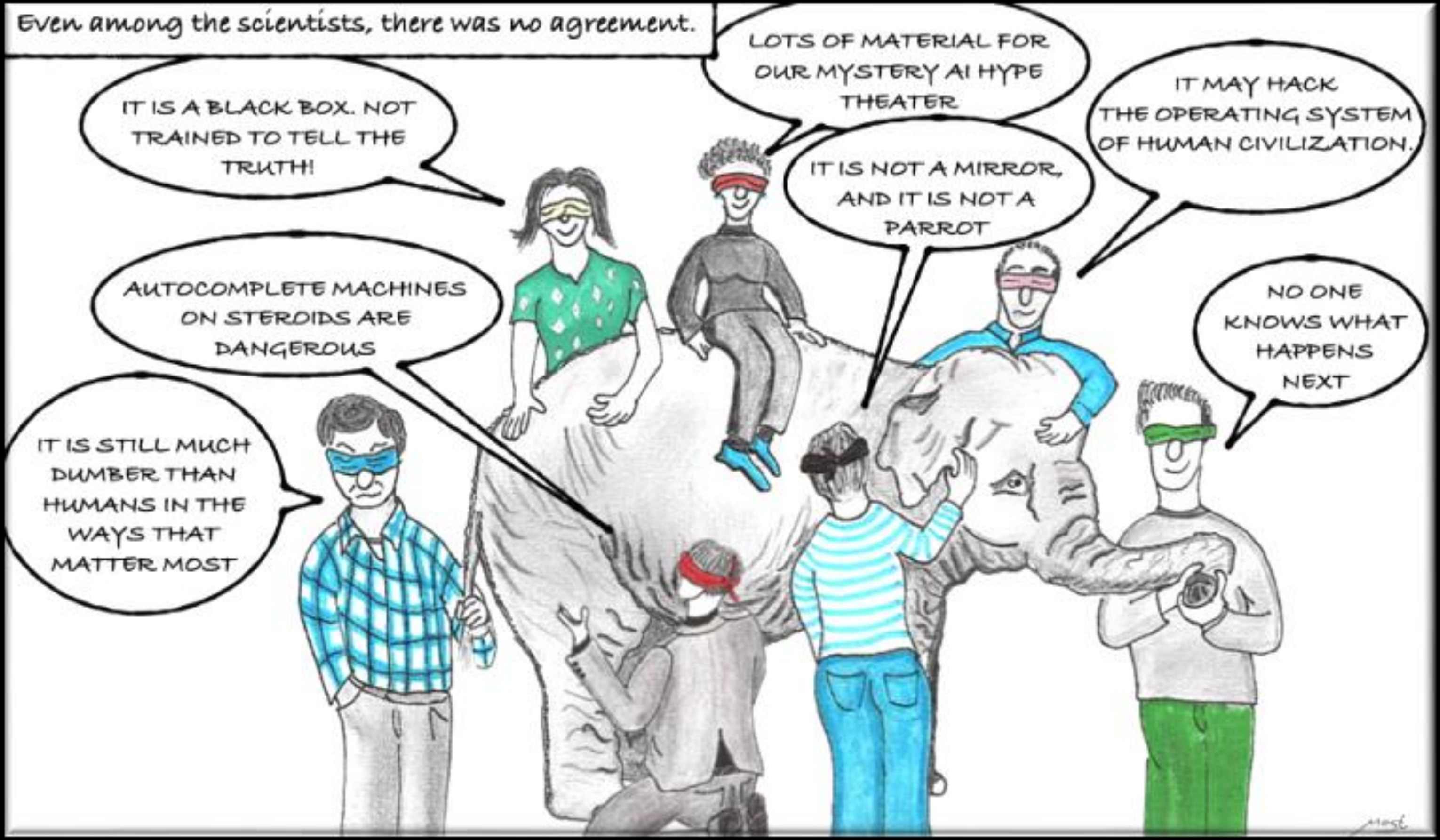
First published: 04 July 2023 | https://doi.org/10.1111/cogs.13309 | Citations: 1



Article

Human-like systematic generalization through a meta-learning neural network

https://doi.org/10.1038/s41586-023-06668-3 Brenden M. Lake & Marco Baroni

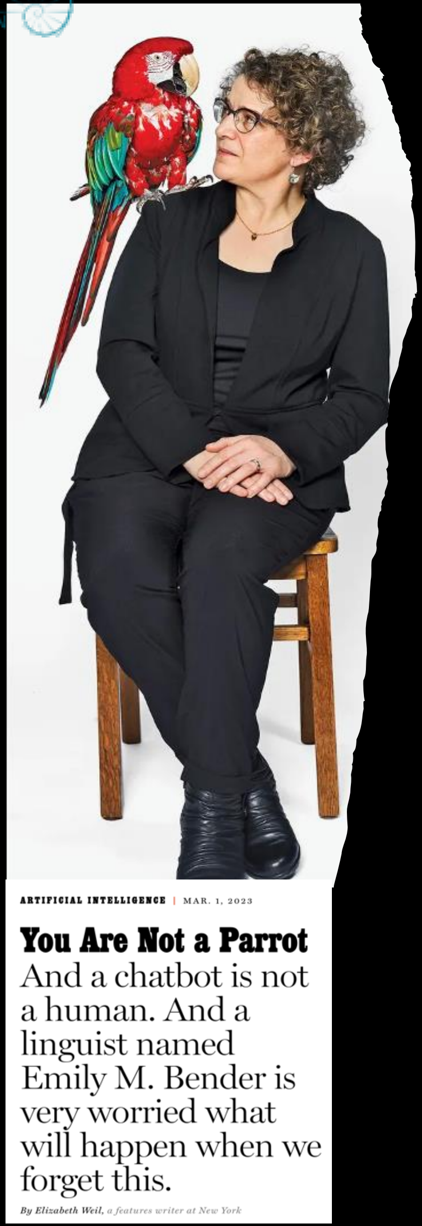


Even among the scientists, there was no agreement.
IT IS A BLACK BOX. NOT TRAINED TO TELL THE TRUTH!
LOTS OF MATERIAL FOR OUR MYSTERY AI HYPE THEATER
IT MAY HACK THE OPERATING SYSTEM OF HUMAN CIVILIZATION.
IT IS NOT A MIRROR, AND IT IS NOT A PARROT
AUTOCOMPLETE MACHINES ON STEROIDS ARE DANGEROUS
NO ONE KNOWS WHAT HAPPENS NEXT
IT IS STILL MUCH DUMBER THAN HUMANS IN THE WAYS THAT MATTER MOST

Artist: Moritz Strasser

(Agrawal et al., 2023; y Arcas, 2022; Lake & Baroni, 2023; Strasser & Strasser, 2024; Trott et al., 2023)

Landscape of opinions about LLMs

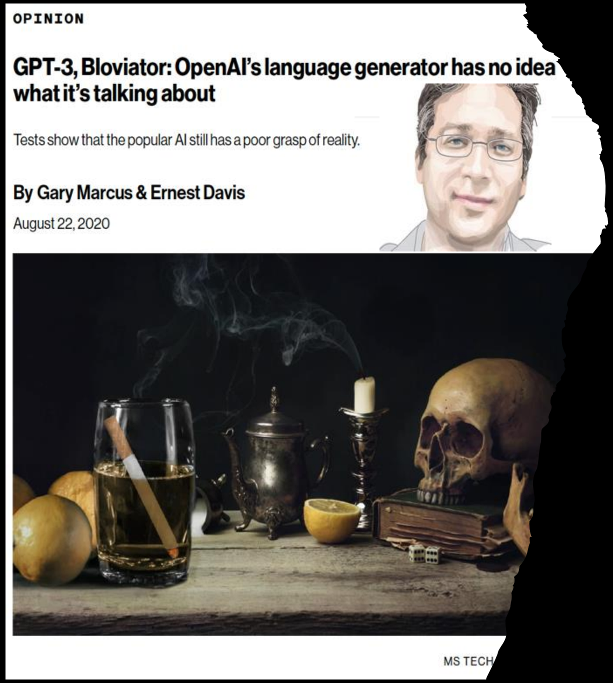


ARTIFICIAL INTELLIGENCE | MAR. 1, 2023

You Are Not a Parrot

And a chatbot is not a human. And a linguist named Emily M. Bender is very worried what will happen when we forget this.

By Elizabeth Weil, *a features writer at New York*



OPINION

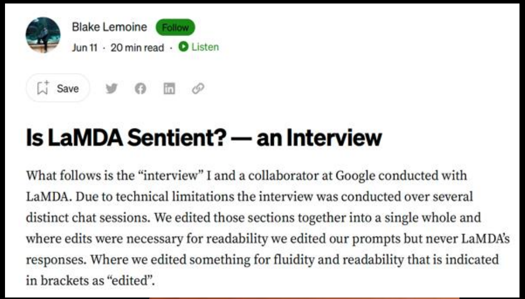
GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about

Tests show that the popular AI still has a poor grasp of reality.

By Gary Marcus & Ernest Davis

August 22, 2020

MS TECH



Blake Lemoine Follow

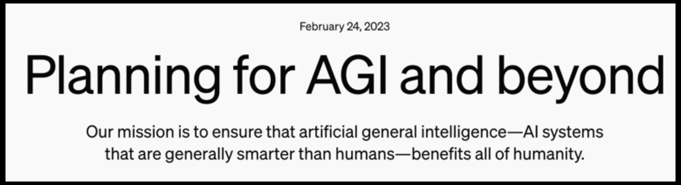
Jun 11 · 20 min read · Listen

Save

Is LaMDA Sentient? — an Interview

What follows is the "interview" I and a collaborator at Google conducted with LaMDA. Due to technical limitations the interview was conducted over several distinct chat sessions. We edited those sections together into a single whole and where edits were necessary for readability we edited our prompts but never LaMDA's responses. Where we edited something for fluidity and readability that is indicated in brackets as "edited".

OpenAI



February 24, 2023

Planning for AGI and beyond

Our mission is to ensure that artificial general intelligence—AI systems that are generally smarter than humans—benefits all of humanity.

(Bender et al., 2021; Lemoine, 2022; Marcus & Davis, 2020; Open-AI; Weil, 2023)

Social Robotics

APPROACHES EXPLORING THE EVENTUAL SOCIAL STATUS WE ATTRIBUTE TO SOCIAL ROBOTS

Many studies in HRI have shown that humans do not only attribute agency but also social skills to robots.

Kerstin Dautenhahn (2007)

- ❖ examined different paradigms regarding ‘social relationships’ of robots and people interacting with them. Taking social and interactive skills of robots as a necessary requirement for the success of many human-robot interactions (HRIs) she discussed the nature of interactivity and ‘social behavior’.

Johanna Seibt et al. (2020)

- ❖ *'sociomorphing'*
  perception of actual non-human social capacities as a form of sense-making of a social other (not anthropomorphizing!) and their phenomenological counterparts 'types of experienced sociality’ to relate robotic properties to types of human experiences and interactive dispositions

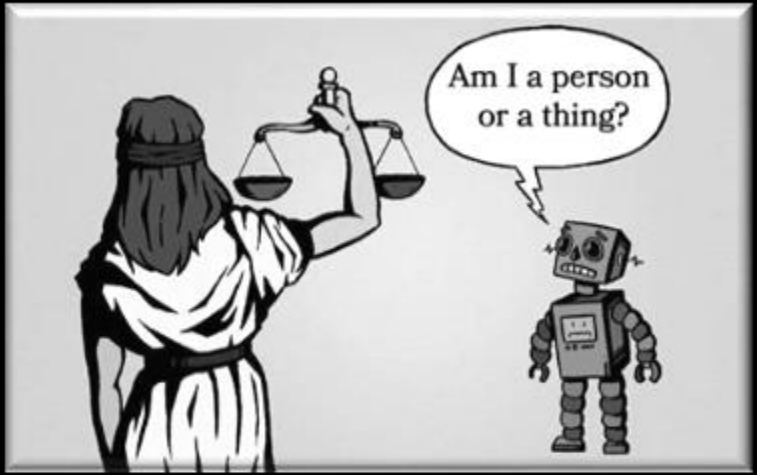
The application of generative AI in social robotics will give rise to many new debates and studies.

(Dautenhahn, 2007; Seibt et al., 2020)

Insights I gained concerning debates about LLMs

WHAT DO WE DO WHEN WE INTERACT WITH LLMS?

WE CANNOT REDUCE ALL OF OUR INTERACTIONS WITH LLMS (AND ESPECIALLY WITH FUTURE PRODUCTS OF GENERATIVE AI) TO MERE TOOL USE



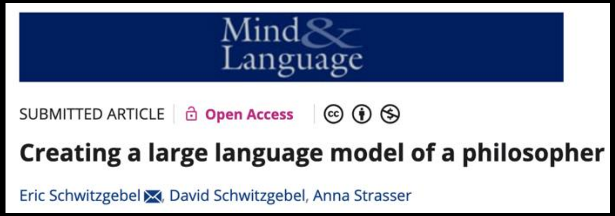
AI systems increasingly occupy a middle ground between genuine personhood and mere causally describable machines.

➢ certain artificial systems are neither persons nor things

❖ **BUT there is no philosophical terminology to describe what they are instead**

→ Rethink conceptual frameworks, which so clearly distinguish between tools as inanimate, asocial things and humans as social, rational, and moral interaction partners!

Lessons learned with LLMs

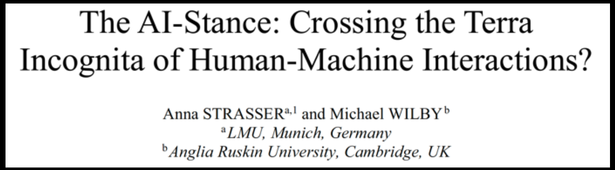


Mind & Language

SUBMITTED ARTICLE | Open Access

Creating a large language model of a philosopher

Eric Schwitzgebel, David Schwitzgebel, Anna Strasser



The AI-Stance: Crossing the Terra Incognita of Human-Machine Interactions?

Anna STRASSER[a,1] and Michael WILBY[b]
[a] LMU, Munich, Germany
[b] Anglia Ruskin University, Cambridge, UK

A HUMAN-MADE BOOK IN THE AGE OF MACHINE-GENERATED TEXTS

Anna Strasser (Ed.)
Anna's AI Anthology
How to live with smart machines?
MoMo berlin 9
THE HUMAN & THE AI
MADE by HUMANS

Anna´s AI Anthology

How to live with smart machines?

With the release of ChatGPT, large language models (LLMs) have become a prominent topic of international public and scientific debate.

The genie is out of the bottle, but does it have a mind?

Can philosophical considerations help us to work out how we can live with such smart machines? In this book, distinguished philosophers explore questions such as whether these new machines are able to act, whether they are social agents, whether they have communicative skills, and if they might even become conscious.

The book includes contributions from:

Syed AbuMusab, Daniel Dennett, Frederic Gilbert, Joshua Rust, Anna Strasser, Constant Bonard, Paula Droege, Ying-Tung Lin, Eric Schwitzgebel, Alessio Tacca, Stephen Butterfill, Keith Frankish, Sven Nyholm, Henry Shevlin, Michael Wilby

As a bonus, the book contains a 44 page, colored graphic novel by Anna & Moritz Strasser.

xenomoi verlag

Humans and Smart Machines as Partners in Thought?

A hybrid workshop about large language models

UC RIVERSIDE

hosted by the UC Riverside Philosophy Department

- organized by Anna Strasser & Eric Schwitzgebel
- supported by

INTERNATIONAL SOCIAL ONTOLOGY SOCIETY

Center for Ideas and Society

SAVE THE DATE

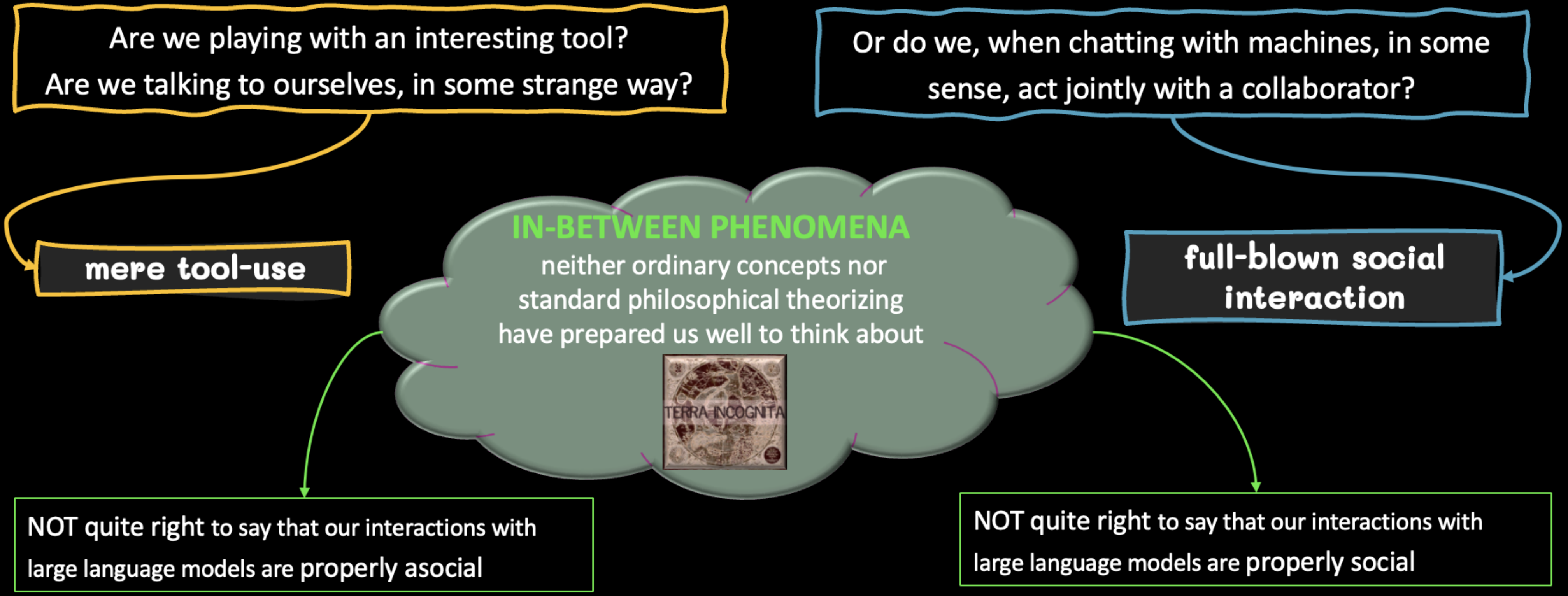
10-11 MAY 2023

Advocate a thorough, gradual approach describing a multi-dimensional spectrum of all kinds of social interactions

(Schwitzgebel et al., 2023; Strasser, 2024; Strasser et al., 2023; Strasser & Wilby, 2023; Strasser & Schwitzgebel, 2024)

Quasi-sociality

WHAT DO WE DO WHEN WE INTERACT WITH LLMS?



Are we playing with an interesting tool?
Are we talking to ourselves, in some strange way?

Or do we, when chatting with machines, in some sense, act jointly with a collaborator?

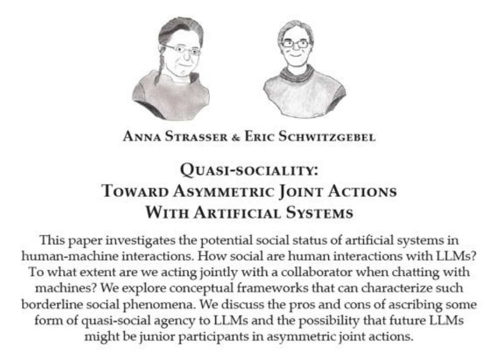
mere tool-use

IN-BETWEEN PHENOMENA
neither ordinary concepts nor standard philosophical theorizing have prepared us well to think about

full-blown social interaction

NOT quite right to say that our interactions with large language models are **properly asocial**

NOT quite right to say that our interactions with large language models are **properly social**



ANNA STRASSER & ERIC SCHWITZGEBEL

QUASI-SOCIALITY:
TOWARD ASYMMETRIC JOINT ACTIONS
WITH ARTIFICIAL SYSTEMS

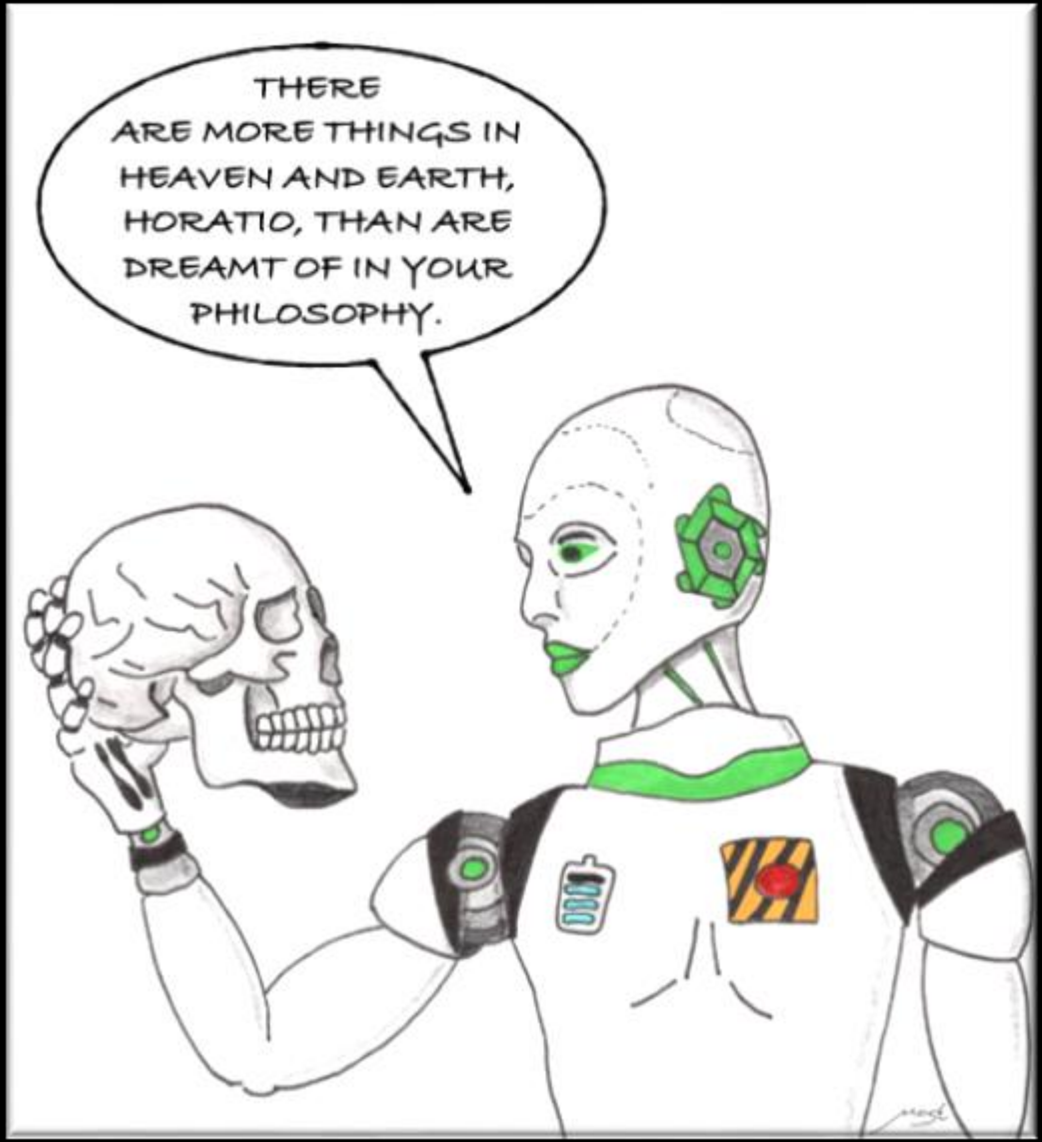
This paper investigates the potential social status of artificial systems in human-machine interactions. How social are human interactions with LLMs? To what extent are we acting jointly with a collaborator when chatting with machines? We explore conceptual frameworks that can characterize such borderline social phenomena. We discuss the pros and cons of ascribing some form of quasi-social agency to LLMs and the possibility that future LLMs might be junior participants in asymmetric joint actions.

INTERACTIONS WITH LLMS, OR OTHER RECENT AND EMERGING AI SYSTEMS,
ARE, OR CAN BE, QUASI-SOCIAL

- drawing on the human agent's social skills and attributions, that isn't just entirely fictional or pointless
- machine partner can be an entity that rightly draws social reactions and attributions in virtue of having features that make such reactions and attributions more than just metaphorically apt

(Strasser & Schwitzgebel, 2024)

The Terra Incognita – the INBETWEEN

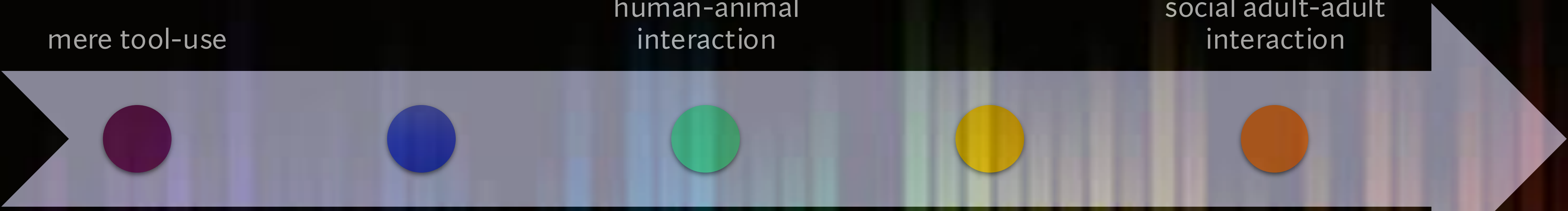


THERE ARE MORE THINGS IN HEAVEN AND EARTH, HORATIO, THAN ARE DREAMT OF IN YOUR PHILOSOPHY.

Artist: Moritz Strasser

(Shakespeare, 2021; Strasser & Strasser, 2024)

A gradual approach



mere tool-use

quasi-social human-machine interaction

quasi-social human-animal interaction

quasi-social adult-infant interaction

social adult-adult interaction

SINGLE-SIDED SOCIALITY

- sociality tossed into a void
- application of social skills
- reactions toward entities who are in no respect social partners, with no capacity for social uptake

QUASI-SOCIALITY

- machines designed in a way that exploits the fact that you will react to it as a social agent; and you, in turn, can exploit that fact about it

[junior partner]

- lifted or scaffolded into complex joint action by the engagement & structuring of the more knowledgeable partner

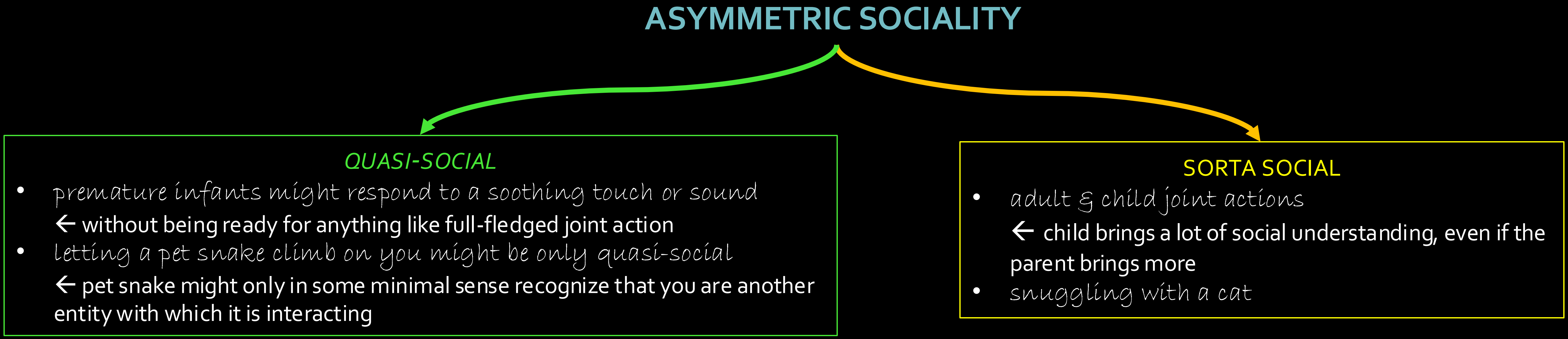
[senior partner]

- knows that they know what the other knows
- fully appreciates the social structure of the interaction they are having

FULL-BLOWN, INTELLECTUALLY DEMANDING, COOPERATIVE SOCIAL INTERACTION

- both partners make second-order mental state attributions and satisfy various other conditions are required for full-blown adult human cooperative action

A spectrum of asymmetric sociality



ASYMMETRIC SOCIALITY

QUASI-SOCIAL

- *premature infants might respond to a soothing touch or sound*
  ← without being ready for anything like full-fledged joint action
- *letting a pet snake climb on you might be only quasi-social*
  ← pet snake might only in some minimal sense recognize that you are another entity with which it is interacting

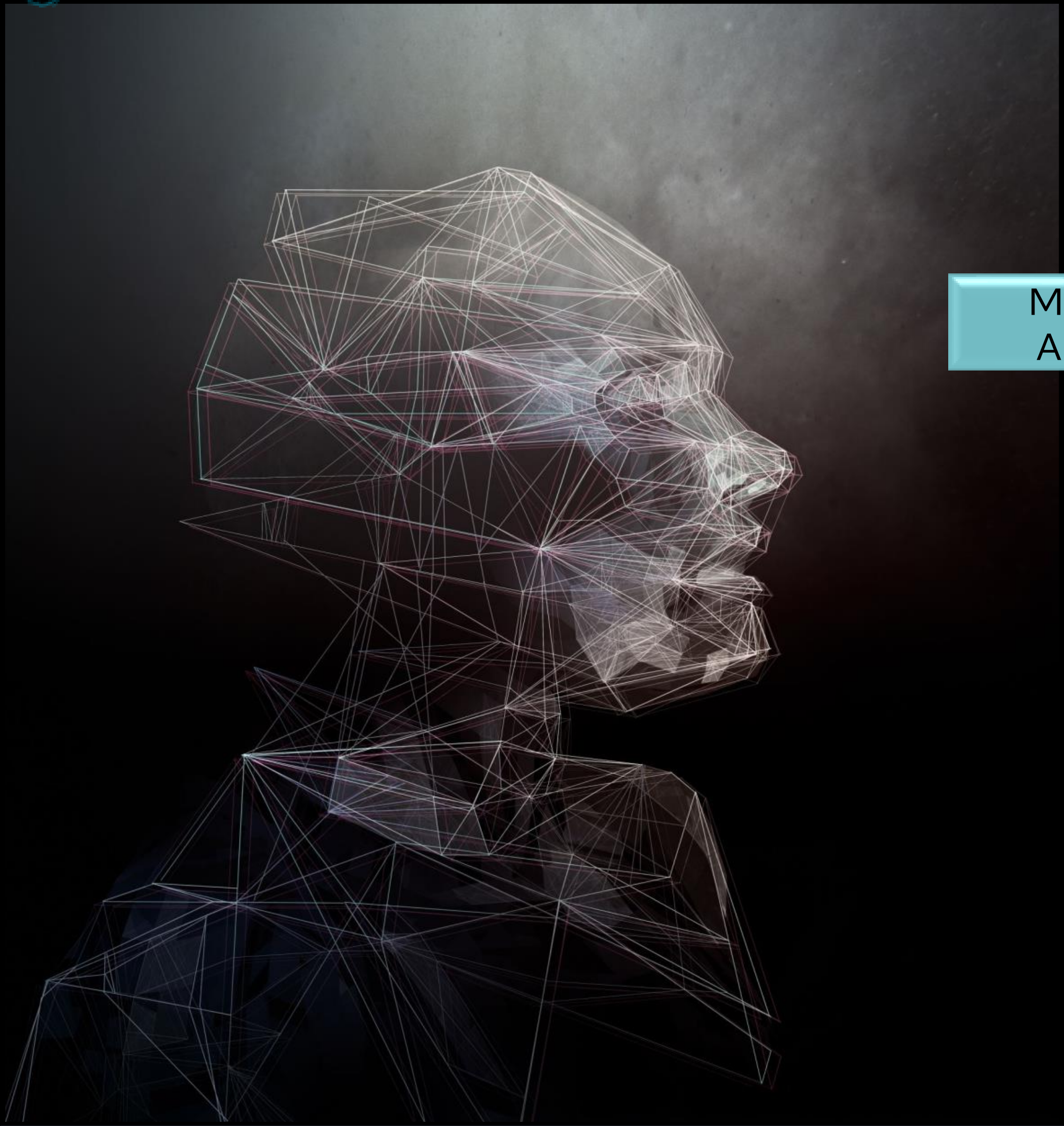
SORTA SOCIAL

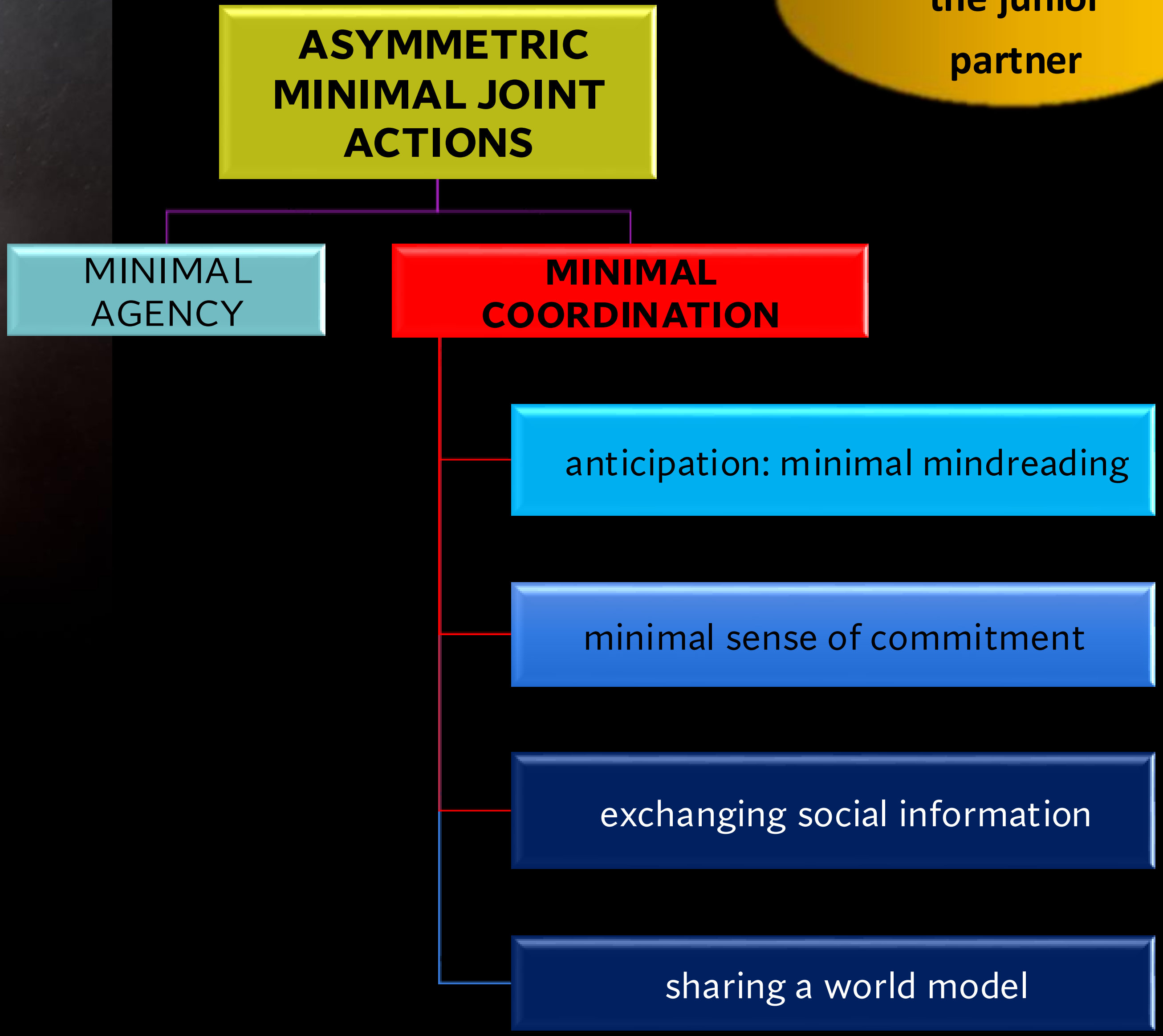
- *adult & child joint actions*
  ← child brings a lot of social understanding, even if the parent brings more
- *snuggling with a cat*

QUASI-SOCIAL INTERACTIONS ARE INTERACTIONS BETWEEN A FULLY SOCIAL AGENT AND SOME PARTNER – WHETHER HUMAN, MACHINE, OR ANIMAL – THAT IS NOT COGNITIVELY CAPABLE OF FULL-FLEDGED SOCIAL JOINT ACTION BUT THAT DOES RESPOND IN A WAY THAT PRODUCTIVELY INVITES FURTHER SOCIAL RESPONSES FROM THE SOCIAL PARTNER.

Asymmetric joint actions

conditions for the junior partner





ASYMMETRIC MINIMAL JOINT ACTIONS

MINIMAL AGENCY

MINIMAL COORDINATION

anticipation: minimal mindreading

minimal sense of commitment

exchanging social information

sharing a world model

Tool kit ‘minimal approaches’

How to conceptualize phenomena in the field of developmental psychology & animal cognition that fall through the sophisticated conceptual net of philosophy

- ❖ questioning the necessity of far too demanding conditions
- ❖ considering multiple realizations of capacities that seemed to be restricted to sophisticated adult humans

Butterfill & Apperly (2013): minimal mindreading | Michael et al. (2016): minimal sense of Commitment | Pacherie (2013): shared intention lite | Strasser (2006): minimal action

Moving from the two-dimensional space to the three-dimensional space

increasing INDISTINGUISHABILITY between machine-generated & human-created text

LLMs live NOT in our social, physical world

LLMs are not embodied

But they may play a role in our world of language games.

explore the to-be-expected implications of the experience that our sociality gains traction within communicative exchanges in HRI

IN-DISTINGUISH-ABILITY

- so far we can easily distinguish humans from robots

EMBODIMENT

- cause changes in our physical world

SHARING WORLD MODELS

- ?

All this would not have been possible if I had not interacted with people & machines

Daniel Dennett

Eric Schwitzgebel

Mathew Crosby

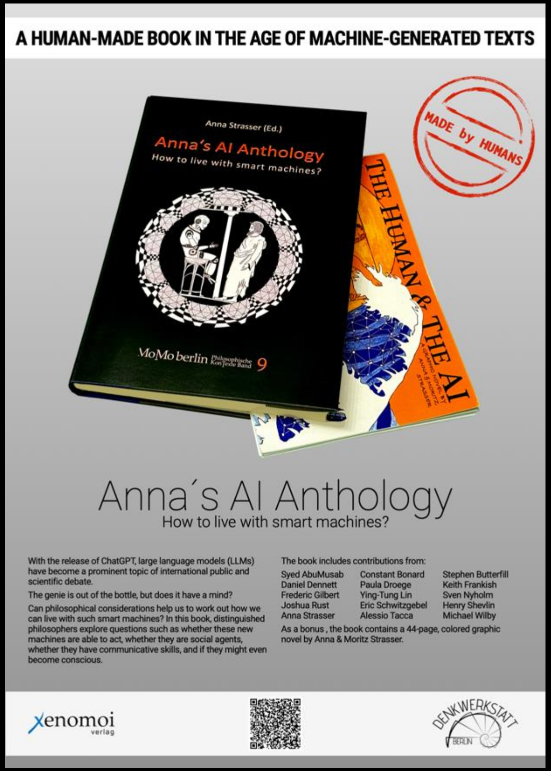
David Schwitzgebel

Mike Wilby

DigiDan

In case you want to order Anna's AI Anthology

Thank you !



A HUMAN-MADE BOOK IN THE AGE OF MACHINE-GENERATED TEXTS

Anna Strasser (Ed.)
Anna's AI Anthology
How to live with smart machines?
MoMo berlin Philosophische KonTexte Band 9
THE HUMAN & THE AI
A GRAPHIC NOVEL BY ANNA & MORITZ STRASSER
MADE by HUMANS

Anna´s AI Anthology

How to live with smart machines?

With the release of ChatGPT, large language models (LLMs) have become a prominent topic of international public and scientific debate.

The genie is out of the bottle, but does it have a mind?

Can philosophical considerations help us to work out how we can live with such smart machines? In this book, distinguished philosophers explore questions such as whether these new machines are able to act, whether they are social agents, whether they have communicative skills, and if they might even become conscious.

The book includes contributions from:

Syed AbuMusab
Daniel Dennett
Frederic Gilbert
Joshua Rust
Anna Strasser
Constant Bonard
Paula Droege
Ying-Tung Lin
Eric Schwitzgebel
Alessio Tacca
Stephen Butterfill
Keith Frankish
Sven Nyholm
Henry Shevlin
Michael Wilby

As a bonus , the book contains a 44-page, colored graphic novel by Anna & Moritz Strasser.

xenomoi verlag

DENKWERKSTATT BERLIN

References

Agrawal, A., Mackey, L., & Kalai, A. T. (2023). *Do Language Models Know When They're Hallucinating References?* (arXiv:2305.18248). arXiv. http://arxiv.org/abs/2305.18248

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Butterfill, S. A., & Apperly, I. A. (2013). How to Construct a Minimal Theory of Mind. *Mind & Language*, *28*(5), 606–637. https://doi.org/10.1111/mila.12036

Dautenhahn, K. (2007). Socially intelligent robots: Dimensions of human–robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 679–704. https://doi.org/10.1098/rstb.2006.2004

Lake, B. M., & Baroni, M. (2023). Human-like Systematic Generalization through a Meta-learning Neural Network. *Nature*, 1–7. https://doi.org/10.1038/s41586-023-06668-3

Lemoine, B. (2022, June 11). Is LaMDA Sentient? — An Interview. *Medium*. https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917

Marcus, G., & Davis, E. (2020). GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. *MIT Technology Review*. https://www.technologyreview. com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion

Michael, J., Sebanz, N., & Knoblich, G. (2016). The Sense of Commitment: A Minimal Approach. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01968

Open-AI. (n.d.). *Planning for AGI and beyond*. Retrieved 6 May 2024, from https://openai.com/index/planning-for-agi-and-beyond

Pacherie, E. (2013). Intentional joint agency: Shared intention lite. *Synthese*, *190*(10), 1817–1839. https://doi.org/10.1007/s11229-013-0263-7

References

Schwitzgebel, E., Schwitzgebel, D., & Strasser, A. (2023). Creating a large language model of a philosopher. *Mind & Language*, 1–23. https://doi.org/10.1111/mila.12466

Seibt, J., Vestergaard, C., & Damholdt, M. F. (2020). Sociomorphing, Not Anthropomorphizing: Robophilosophy 2020. *Culturally Sustainable Social Robotics*, 51–67. https://doi.org/10.3233/FAIA200900

Shakespeare, W. (2021). *Complete works* (R. Proudfoot, A. Thompson, D. S. Kastan, & H. R. Woudhuysen, Eds.). The Arden Shakespeare.

Strasser, A. (2006). *Kognition künstlicher Systeme*. De Gruyter. https://doi.org/10.1515/9783110321104

Strasser, A. (Ed.). (2024). *Anna's AI Anthology. How to live with smart machines?* xenomoi Verlag.

Strasser, A., Crosby, M., & Schwitzgebel, E. (2023). How Far Can We Get in Creating a Digital Replica of a Philosopher? In R. Hakli, P. Mäkelä, & J. Seibt (Eds.), *Social Robots in Social Institutions* (pp. 371–380). IOS Press. https://doi.org/10.3233/FAIA220637

Strasser, A., & Schwitzgebel, E. (2024). Quasi-sociality: Toward Asymmetric Joint Actions. In *Anna's AI Anthology. How to live with smart machines?* xenomoi Verlag.

Strasser, A., & Wilby, M. (2023). The AI-Stance: Crossing the Terra Incognita of Human-Machine Interactions? In *Social Robots in Social Institutions* (pp. 286–295). IOS Press. https://doi.org/10.3233/FAIA220628

Trott, S., Jones, C., Chang, T., Michaelov, J., & Bergen, B. (2023). Do Large Language Models Know What Humans Know? *Cognitive Science*, *47*(7), e13309. https://doi.org/10.1111/cogs.13309

Weil, E. (2023, March 1). *You Are Not a Parrot*. Intelligencer. https://nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html

y Arcas, B. A. (2022). Do Large Language Models Understand Us? *Daedalus*, *151*(2), 183–197. https://doi.org/10.1162/daed_a_01909