INBETWEEN
Tools and Social Agents.
On the Status of LLMs in Human-Machine Interactions

Invited session: AI for Philosophers: TOOL or AGENT?

XXV World Congress of Philosophy

Philosophy across Boundaries

Sapienza University of Rome, August 1-8, 2024

Anna Strasser, DenkWerkstatt Berlin

INTRODUCTION

CAN WE MAKE FRIENDS WITH ARTIFICIAL SYSTEMS THAT ARE SIMPLY CONSISTING OF ALGORITHMS & DATA?

Is this deeply unsettling?

IF interactions with software would be the most meaningful and important social interactions one has.

MOTHERBOARD
TECH BY VICE

'It's Hurting Like Hell': AI Companion Users Are In Crisis, Reporting Sudden Sexual Rejection

Replika, the "AI companion who cares," has undergone some abrupt changes to its erotic roleplay features, leaving many users confused and heartbroken.

By Samantha Cole

- 2023

Replika users feel like losing their best friend after an update

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".

- 2022

Blake Lemoine claimed that Lambda had consciousness & sentience

- 2018

Akihiko Kondo married his beloved waifu, a hologram

(Cole, 2023; Dooley & Ueno, 2022; Lemoine, 2022)

Controversial debate

ATTRIBUTION OF KNOWLEDGE | UNDERSTANDING | SYSTEMATIC GENERALIZATION …

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.

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

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

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.

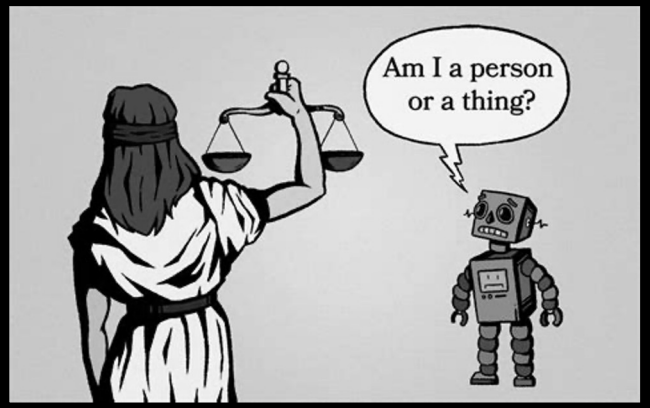
(Agrawal et al., 2023; y Arcas, 2022; Lake & Baroni, 2023; Strasser & Strasser, 2024; Trott et al., 2023)
(Bender et al., 2021; Open-AI, n.d ; Heaven, 2020; Marcus & Davis, 2020; Weil, 2023

My question & main claim

What are we doing when we interact with LLMs?

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

❖ AI systems increasingly occupy in HMIs a middle ground between genuine personhood & mere causally describable machines



- Is an LLM or a robot developed with generative AI technology a person or a thing? → neither nor

BUT so far we have no philosophical terminology to describe what it is instead!

→ rethink our conceptual framework, which so clearly distinguishes between tools as inanimate things and humans as social, rational, and moral interaction partners

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

(Strasser & Schwitzgebel 2024)

THE INBETWEEN

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

full-blown social interaction

IN-BETWEEN PHENOMENA

neither ordinary concepts nor standard philosophical theorizing have prepared us well to think about them

1 expand concept of tool-use
(add complex tools with social features)

2 expand conception of social interactions
(add non-living social agents)

3 *add a third category*

Search for a gradual conceptual framework
(question the dichotomy)

All routes are full of construction sides!

…, I invite you to join me to find a way through the jungle of the Terra Incognita.

1

Emphasize the differences between humans & machines

- LLMs are in their causal genesis functionally (neurobiologically & cognitively) absolutely dissimilar to an intelligent, sentient human being

BUT

difficult to argue for potential multiple realizations of socio-cognitive capacities that are normally only ascribed to living agents

2

Argue for similarities between humans & machines

- In immediate interactions, the AI seems functionally (i.e., conversationally) similar to an intelligent, sentient human being (Lemoine, 2022)

BUT

wrongly overemphasize similarities between humans and machines

3

The problem of conceptualizing the INBETWEEN does not disappear if we introduce another category.

➢ If we establish a conceptual framework that contains three categories, we will then have two in-betweens that we cannot conceptualize

Motivations (I)

Philosophy poses too demanding conditions

too demanding conditions

- describing ideal cases that are rarely found in everyday life

→ abilities of children, non-human animals , and artificial systems fall through the conceptual net

sophisticated terminology of philosophy prevents us from grasping the in-between
→ conceptual frameworks that can distinguish more finely-grained instances across a wider spectrum

- capture phenomena one finds in developmental psychology, animal cognition, and AI

thinking about how to conceptualize the INBETWEEN by discussing notions like

- quasi-social versus full-fledged social
- minimal agency versus full-fledged agency
- asymmetric quasi-social joint actions versus full-fledged joint actions

(Strasser, 2006, 2008)

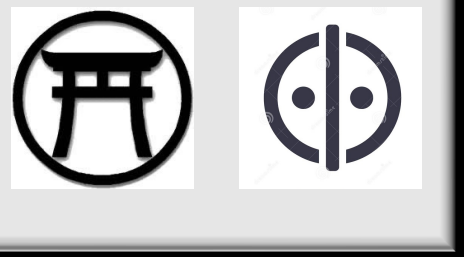
Motivations (II)

QUESTIONING THE DICHOTOMY BETWEEN ANIMATE AND INANIMATE

1

Western conception is just one conception of many

shintoism & animism



artificially constructed dichotomies

2

global rights-of-nature movement

rivers in India & New Zealand, & Canada were granted legal personhood

- legal steps linking Western & Indigenous worldviews
- first step towards promoting a kinship-oriented worldview (Salmón, 2000)

legal personhood for non-living entities

3

notion of a social agent has proven to be changeable

e.g. status of women, children, other ethnicities, non-human animals

scope of sociality can be expanded

(Gunkel, 2023; Jensen & Blok, 2013; Robertson, 2014, 2017 | O'Donnell & Talbot-Jones, 2017; Bunten et al., 2021)

Motivations (III)

QUESTIONING THE DICHOTOMY BETWEEN ANIMATE AND INANIMATE

4

Similarities with human-human interactions

- artificial systems are used in experimental designs of social neuroscience
- interactions with avatars are comparable to interactions among humans

→ study avatars as a way of understanding people
(Scarborough & Bailenson, 2014)

If interactions with artificial systems would not have any similarities with human-human interactions, we could not use them to explore human behavior.

Motivations from an ethical perspective

QUESTIONING THE DICHOTOMY BETWEEN ANIMATE AND INANIMATE

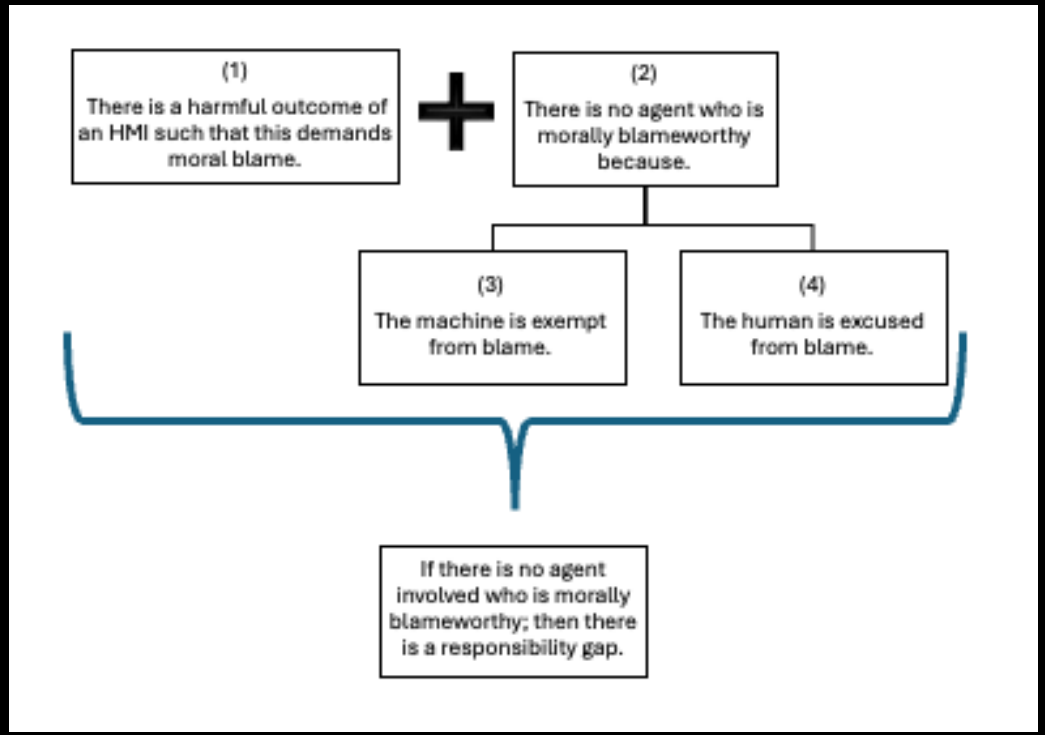
Hard-core instrumental view

NON-LIVING THINGS CAN NEITHER HAVE MORAL AGENCY NOR MORAL PATIENCY

BUT

IF ARTIFICIAL SYSTEMS ARE MERE TOOLS **THEN**

1. question previously justified justifications for HMI in which the human interaction partners were excused
   - because artificial systems are exempt
2. live with many responsibility gaps
   - because humans are excused & artificial systems are exempt
3. difficulties in arguing for social norms guiding our behavior toward artificial systems
   - because artificial systems have no moral patiency



(Strasser, 2020; Wilby & Strasser, 2024)

Motivations from an ethical perspective

QUESTIONING THE DICHOTOMY BETWEEN ANIMATE AND INANIMATE

In expectation of AGI view

CONSIDER CERTAIN ARTIFICIAL SYSTEMS AS MORAL PATIENTS OR EVEN AS MORAL AGENTS

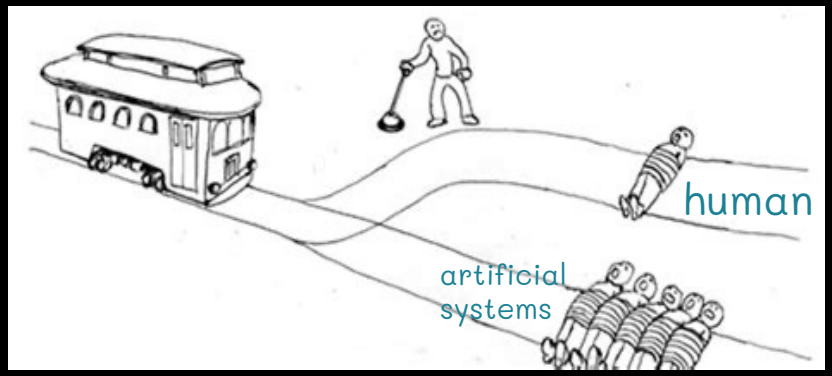
BUT

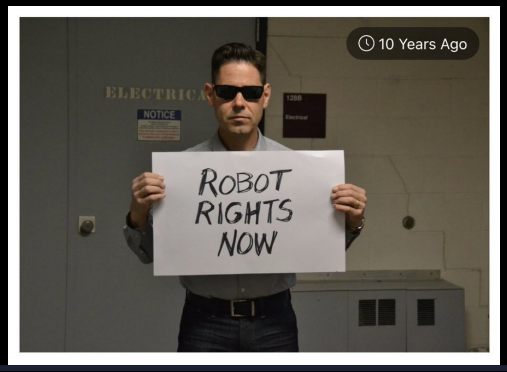
THIS MAY LEAD TO THE IDEA OF ARTIFICIAL LIFE

1. risk of prioritizing artificial agents over human beings
2. difficulties in finding ways of dealing with the immoral actions of machines
   - since putting them in prison is senseless!

less radical position

- risk of over-attributing moral agency and patiency





(Gunkel, 2020; Strasser, forthcoming)

Finding our way through the jungle

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

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

The way through the jungle

QUESTIONING THE DICHOTOMY BETWEEN ANIMATE AND INANIMATE

Hard-core instrumental view

instrumental view
artificial agents cannot be participants in joint actions

In expectation of AGI view

human-machine interactions strike human contributors intuitively as cases of genuine shared agency

→ **MID-WAY POINT BETWEEN**

sub-intentional interactions that amount to ‘mere behavior’ (tool use)

rich, intellectualist views of shared agency

Towards asymmetric joint actions

NO NECESSITY OF AN EQUAL DISTRIBUTION OF ABILITIES AMONG ALL PARTICIPANTS

DEVELOPMENTAL PSYCHOLOGY

- joint action of adults and children
- children = socially interacting beings

ARTIFICIAL INTELLIGENCE

- joint action of human beings & artificial systems
- artificial systems =?= socially interacting entities

ADULT & CHILD

ROBOT & HUMAN
LLM & HUMAN

ASYMMETRIC JOINT ACTIONS

Inbetween mere tool-use and social interactions

TOWARDS ASYMMETRIC JOINT ACTIONS

ASYMMETRIC MINIMAL JOINT ACTIONS

- **MINIMAL AGENCY**
- MINIMAL COORDINATION
  - anticipation: minimal mindreading
  - minimal sense of commitment
  - sharing a world model | exchanging social information

Conclusion

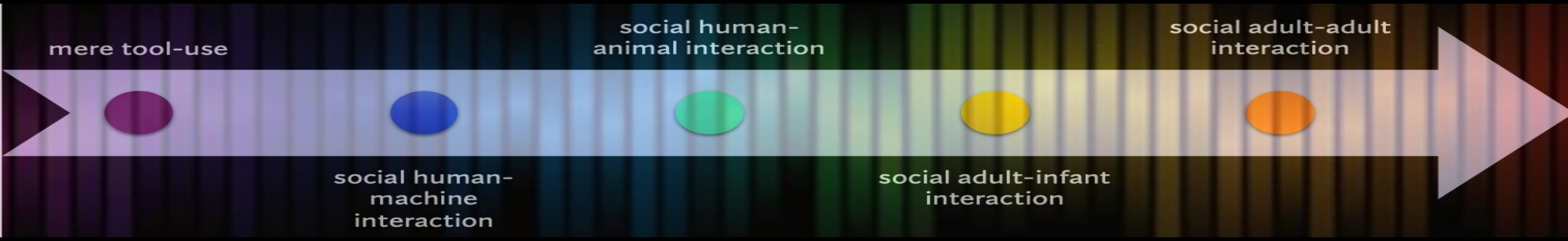
Consider the possibility of questioning the dichotomy between animate and inanimate entities

After all, we might be confronted with a new game.

The main aim of this talk was to prepare the grounds for questioning the dichotomy between animate and inanimate entities, as this is an important presupposition for developing a conceptual framework that can capture phenomena that I locate in the INBETWEEN.

If I am successful with this, I can argue for a gradual approach describing all kinds of social interactions, and finally answer the question of what we are doing when we interact with LLMs— what status artificial systems have in HMIs.

Then we can stop reducing all our interactions with artificial systems (and especially with future products of generative AI) to mere tool use.



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

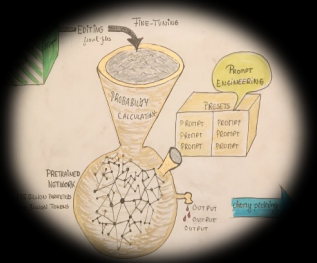
Daniel Dennett

Eric Schwitzgebel

Mathew Crosby

David Schwitzgebel

Mike Wilby



DigiDan

Thank you !

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

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

Barkham, P. (2021, July 25). Should rivers have the same rights as people? *The Guardian*. https://www.theguardian.com/environment/2021/jul/25/rivers-around-the-world-rivers-are-gaining-the-same-legal-rights-as-people

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

Bunten, A., Iorns, C., Townsend, J., & Borrows, L. (2021, June 3). *Rights for nature: How granting a river 'personhood' could help protect it*. The Conversation. http://theconversation.com/rights-for-nature-how-granting-a-river-personhood-could-help-protect-it-157117

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

Cole, S. (2023). 'It's Hurting Like Hell': AI Companion Users Are In Crisis, Reporting Sudden Sexual Rejection. *Vice*. https://www.vice.com/en/article/y3py9j/ai-companion-replika-erotic-roleplay-updates

Dooley, B., & Ueno, H. (2022, April 24). This Man Married a Fictional Character. He'd Like You to Hear Him Out. *The New York Times.* https://www.nytimes.com/2022/04/24/business/akihiko-kondo-fictional-character-relationships.html

Gunkel, D. J. (2020). Robot Rights – Thinking the Unthinkable. In *Smart Technologies and Fundamental Rights* (pp. 48–72). Brill. https://doi.org/10.1163/9789004437876_004

---------- (2023). *Person, Thing, Robot: A Moral and Legal Ontology for the 21st Century and Beyond*. The MIT Press. https://doi.org/10.7551/mitpress/14983.001.0001

Henrich, J. P. (2016). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Heyes, C. (2014). False belief in infancy: A fresh look. *Developmental Science*, *17*(5), 647–659. https://doi.org/10.1111/desc.12148

---------- (2015). Animal mindreading: What's the problem? *Psychonomic Bulletin & Review*, *22*(2), 313–327. https://doi.org/10.3758/s13423-014-0704-4

Jensen, C. B., & Blok, A. (2013). Techno-animism in Japan: Shinto Cosmograms, Actor-network Theory, and the Enabling Powers of Non-human Agencies. *Theory, Culture & Society*, *30*(2), 84–115. https://doi.org/10.1177/0263276412456564

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

O'Donnell, E., & Talbot-Jones, J. (2017, March 23). Three rivers are now legally people – but that's just the start of looking after them. *The Conversation*. http://theconversation.com/three-rivers-are-now-legally-people-but-thats-just-the-start-of-looking-after-them-74983

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

References

Perner, J. (1991). *Understanding the representational mind* (pp. xiv, 348). The MIT Press.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526. https://doi.org/10.1017/S0140525X00076512

Robertson, J. (2014). Human Rights vs.Robot Rights: Forecasts from Japan. *Critical Asian Studies*, *46*(4), 571–598. https://doi.org/10.1080/14672715.2014.960707

---------- (2017). *Robo sapiens japanicus: Robots, Gender, Family, and the Japanese Nation*.

Salmón, E. (2000). Kincentric Ecology: Indigenous Perceptions of the Human-Nature Relationship. *Ecological Applications*, *10*(5), 1327–1332. https://doi.org/10.2307/2641288

Scarborough, J. K., & Bailenson, J. N. (2014). Avatar Psychology. In M. Grimshaw (Ed.), *The Oxford Handbook of Virtuality*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199826162.013.033

Sterelny, K. (2012). *The Evolved Apprentice: How Evolution Made Humans Unique*. The MIT Press. https://doi.org/10.7551/mitpress/9780262016797.001.0001

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

---------- (2013). Kognition künstlicher Systeme. In *Kognition künstlicher Systeme*. De Gruyter. https://doi.org/10.1515/9783110321104

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

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

Tomasello, M. (2008). *Origins of human communication* (pp. xiii, 393). MIT Press.

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

Vesper, C., Butterfill, S., Knoblich, G., & Sebanz, N. (2010). A minimal architecture for joint action. *Neural Networks*, *23*(8), 998–1003. https://doi.org/10.1016/j.neunet.2010.06.002

Warneken, F., Chen, F., & Tomasello, M. (2006). Cooperative Activities in Young Children and Chimpanzees. *Child Development*, *77*(3), 640–663. https://doi.org/10.1111/j.1467-8624.2006.00895.x

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

Wilby, M., & Strasser, A. (2024). Situating machines within normative practices: Bridging responsibility gaps with the AI-Stance. In A. Strasser (Ed.), *Anna's AI Anthology. How to live with smart machines?* xenomoi Verlag.