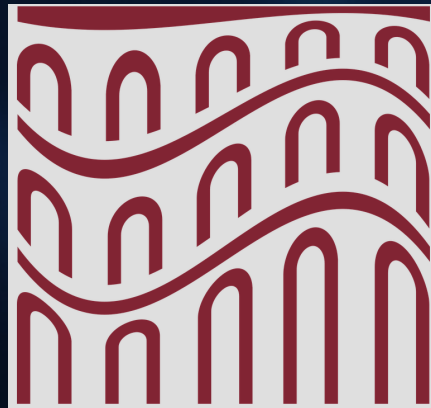

INBETWEENISM

WHY ETHICAL POSITIONS APPEAR OUTDATED IN THE FACE OF THE NEW AI TECHNOLOGY

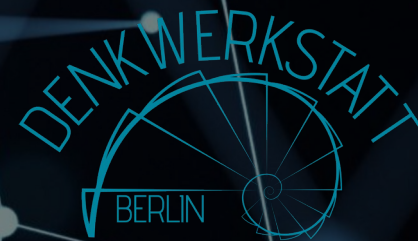
65. Philosophy of Information and Digital Culture



XXV World
Congress of
Philosophy

Philosophy across
Boundaries

Sapienza University of Rome, August 1-8, 2024



Anna Strasser, Denkwerkstatt Berlin

Controversial debate

ATTRIBUTION OF ABILITIES

(COMPREHENSION, KNOWLEDGE, REASONING, AND PHENOMENOLOGICAL CONSCIOUSNESS...)

Many terms that philosophers previously reserved for describing the distinguishing features of humans as rational agents are now being applied to machines, leading to intense debates over such notions as comprehension, knowledge, reasoning, and phenomenological consciousness.

Do Language Models Know When They're Hallucinating References?

Ayush Agrawal
Microsoft Research
t-agrawal@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 Michaelow, 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/s41586-023-06669-3> | Brenden M. Lake¹ & Marco Baroni²

ARTIFICIAL INTELLIGENCE | MAY 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 former 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

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.

Blake Lemoine
Jun 11 · 20 min read · Listen

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".



Implications for ethical approaches

EXPLANATIONS & JUSTIFICATIONS FOR NORM-GUIDED HUMAN BEHAVIOR

interpersonal interactions | interactions with other living beings | interactions with non-living entities

- ❖ CLEAR DEMARCATION BETWEEN LIVING & NON-LIVING ENTITIES
- DICHOTOMY BETWEEN INANIMATE TOOLS & LIVING SOCIAL AGENTS

TOOLS HAVE NEITHER MORAL AGENCY NOR MORAL PATIENCY



debate about the justified ascriptions of all kinds of abilities to products of gen AI



UNCLEAR WHAT ETHICAL STATUS SHOULD BE ASSIGNED TO ARTIFICIAL SYSTEMS IN HMIS

We need a conceptual framework that can capture
INBETWEEN PHENOMENA
to lay the grounds for an ethical approach toward artificial systems.

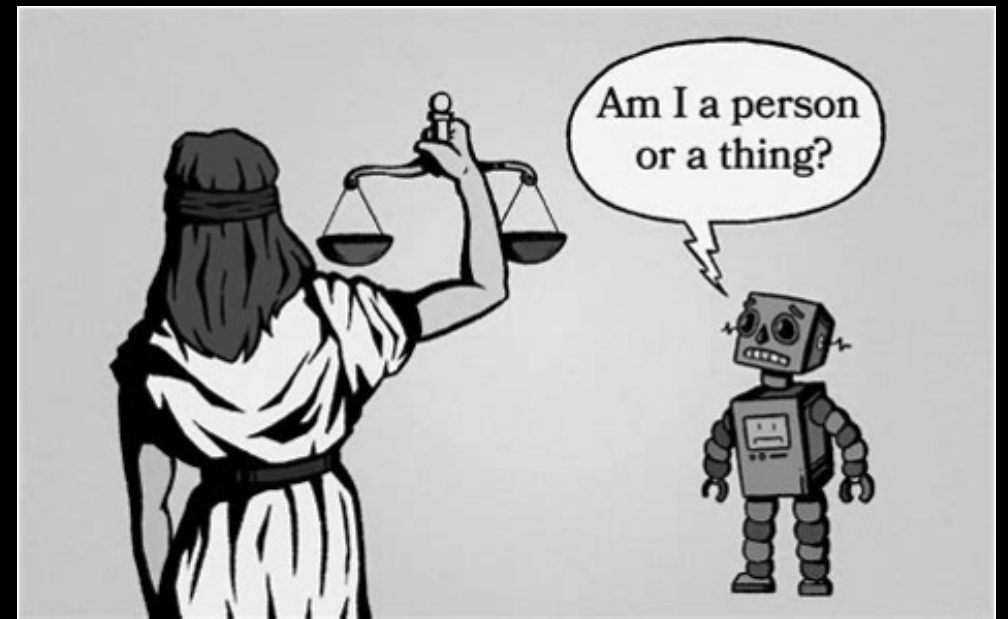
My question & main claim

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

- 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

THE INBETWEEN

WHAT DO WE DO WHEN WE INTERACT WITH LLMs?

Are we playing with an interesting tool?

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

mere tool-use

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

full-blown social
interaction



Routes not taken

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

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

→ emphasize the differences between humans & machines

BUT difficult to argue for potential multiple realizations of socio-cognitive abilities

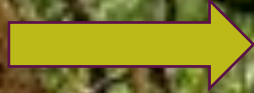
expand concept of social interactions
(add non-living social agents)

→ argue for similarities between humans & machines

BUT wrongly overemphasize similarities between humans and machines

I invite you to join me in finding a way through the jungle of the INBETWEEN.

PHILOSOPHY
POSES TOO
DEMANDING
CONDITIONS



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



sophisticated terminology of philosophy prevents us from grasping the INBETWEEN

→ 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

Motivations

QUESTIONING THE DICHOTOMY BETWEEN ANIMATE AND INANIMATE

1

Western conception is just one conception of many



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



legal personhood for non-living entities

3

Similarities with human-human interactions

- artificial systems are used in experimental designs of social neuroscience
→ study avatars as a way of understanding people



sociality in HMIs can be expanded

Motivations from an ethical perspective

QUESTIONING THE DICHOTOMY BETWEEN ANIMATE AND INANIMATE TO AVOID EXTREME POSITIONS

Hard-core instrumental view

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

all artificial systems should be considered
mere tools
that lack agentive, cognitive, and affective
capacities

→ only human beings (creators or users) can
be held responsible for unfortunate outcomes
of human-machine interactions

In expectation of AGI view

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

certain artificial systems qualify for an ethical
status

→ advocate treating certain artificial systems
analogously to living agents

→ this may, in the end, lead to the idea of
artificial life

Why not remain hard-core instrumentalist?

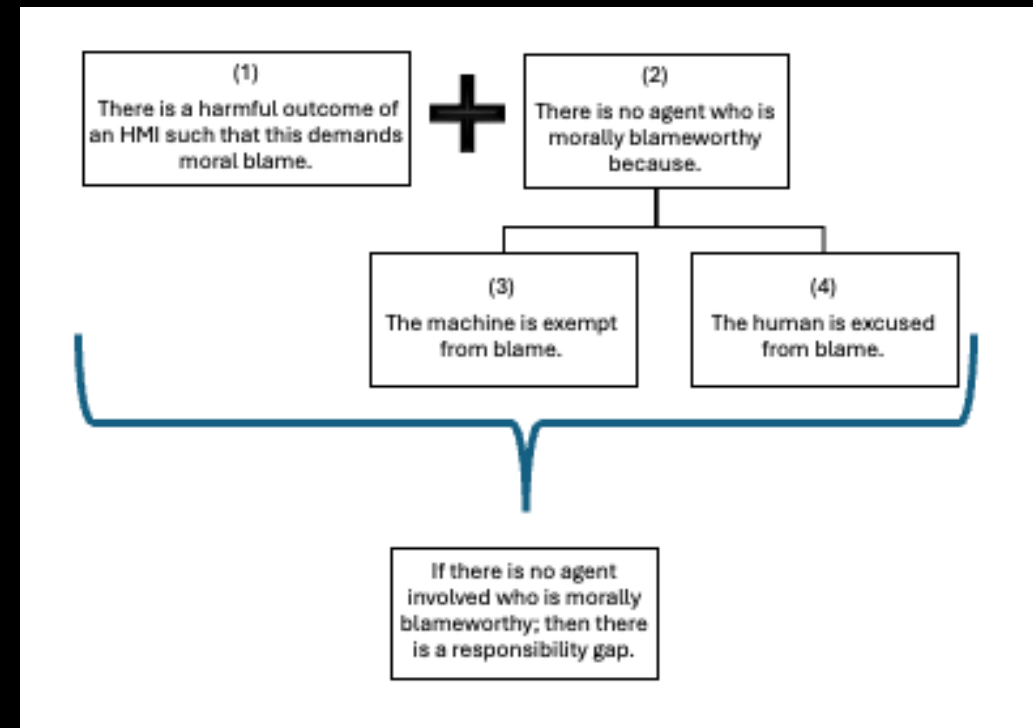
Hard-core instrumental view

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



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 patency



Why not favor the in expectation of AGI view ?

In expectation of AGI view

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

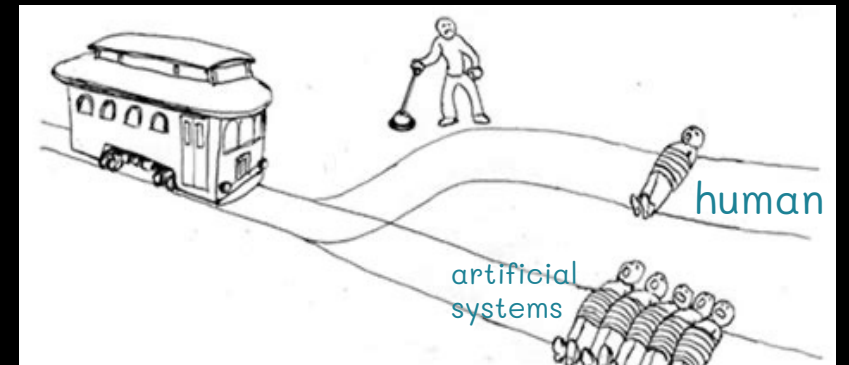


THIS MAY LEAD TO THE IDEA OF ARTIFICIAL LIFE

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



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



The way through the jungle

QUESTIONING THE DICHOTOMY BETWEEN ANIMATE AND INANIMATE

Hard-core instrumental view

instrumental view

artificial systems cannot be participants in social interactions & joint actions

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

In expectation of AGI view

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

rich, intellectualist views of shared agency

NOT mere tool-use

NOT full-blown social interaction

We are not just playing with interesting tools!

Chatting with machines is not like chatting with humans.

IN-BETWEEN PHENOMENA

In order to appreciate the similarities but also the differences, we need a conceptual framework that conceptualizes such in-between phenomena.

Towards asymmetric quasi-social interactions

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

- quasi-social interactions of human beings & artificial systems
- artificial systems =?= quasi-socially interacting entities

ADULT & CHILD

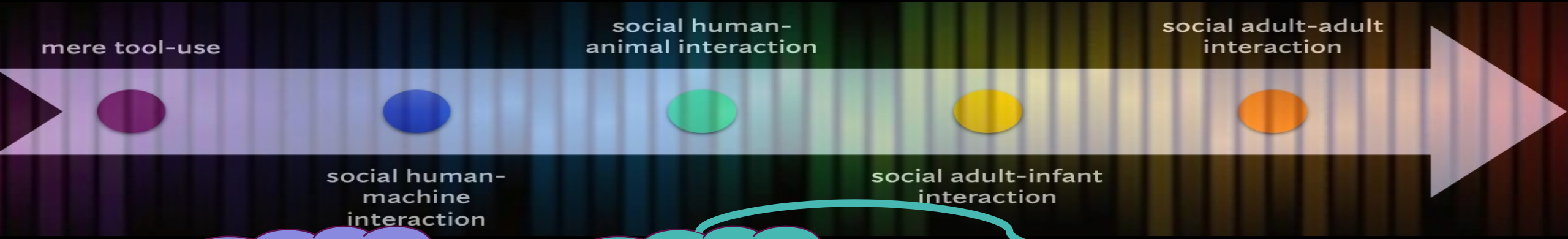


ROBOT & HUMAN
LLM & HUMAN



ASYMMETRIC DISTRIBUTION OF
CONDITIONS

Moving into the spectrum of first quasi-social interactions



QUASI-SOCIALITY

ASYMMETRIC JOINT ACTIONS

conditions for artificial systems

MINIMAL AGENCY

MINIMAL COORDINATION

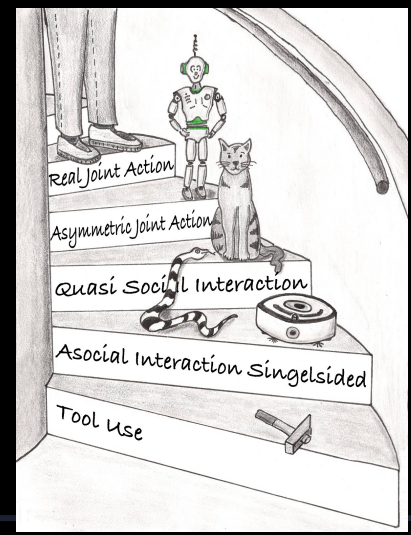
anticipation: minimal mindreading

minimal sense of commitment

sharing a world model | exchanging social information

treating artificial systems as quasi-social partners results in subsequent behavior of the artificial system that feeds back into that way of treating them

- first hints to deny that this is just tool use
- The more their behavior feeds back to this, the more likely it is that their behavior is based on some minimal socio-cognitive abilities.



Conclusion

PREPARE THE GROUNDS FOR QUESTIONING

THE DICHOTOMY BETWEEN ANIMATE AND INANIMATE ENTITIES

**❖ AN IMPORTANT PRESUPPOSITION FOR DEVELOPING A CONCEPTUAL FRAMEWORK THAT
CAN CAPTURE INBETWEEN PHENOMENA**

**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



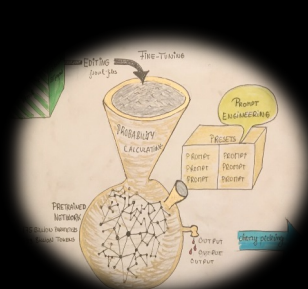
Mathew
Crosby



David
Schwitzgebel



Mike
Wilby



DigiDan

Thank you!



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

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 academic debate. This guide is not of the heftiest, 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 ever become conscious.

The book includes contributions from:

David Aronson	Christoph Menzel	Stephan Ruhrle
David Griesemer	Franka Hees	Walter Frank
David Griesemer	Franka Hees	Walter Frank
David Griesemer	Franka Hees	Walter Frank
David Griesemer	Franka Hees	Walter Frank
David Griesemer	Franka Hees	Walter Frank

As a bonus, the book contains a 48-page colored graphic novel by Anna & Mike Wilby.

xenomoi publishing

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-lambda-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 japonicus: 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.
-

Neither intellectualist nor biological conceptions are wholly convincing

The Intellectualist Approach

attempts to draw a sharp distinction (a “dramatic divide”) between those who are capable of genuine thought and those who aren’t

- have a difficulty explaining how one goes from one side of the divide to the other
- ❖ developmental & comparative psychology suggest that the change is *gradual and not sharp*

The Biological Approach

- attempts to draw the distinction due to a mysterious capacity of our brain to generate consciousness, feeling, subjectivity, and meaning
- fail to explain what the missing quality is, how we can know when it is there and when it is missing
- ❖ why we should suppose that it can only be realized in **electro-chemical brain reactions**, and not in silicon systems, or **neural nets**