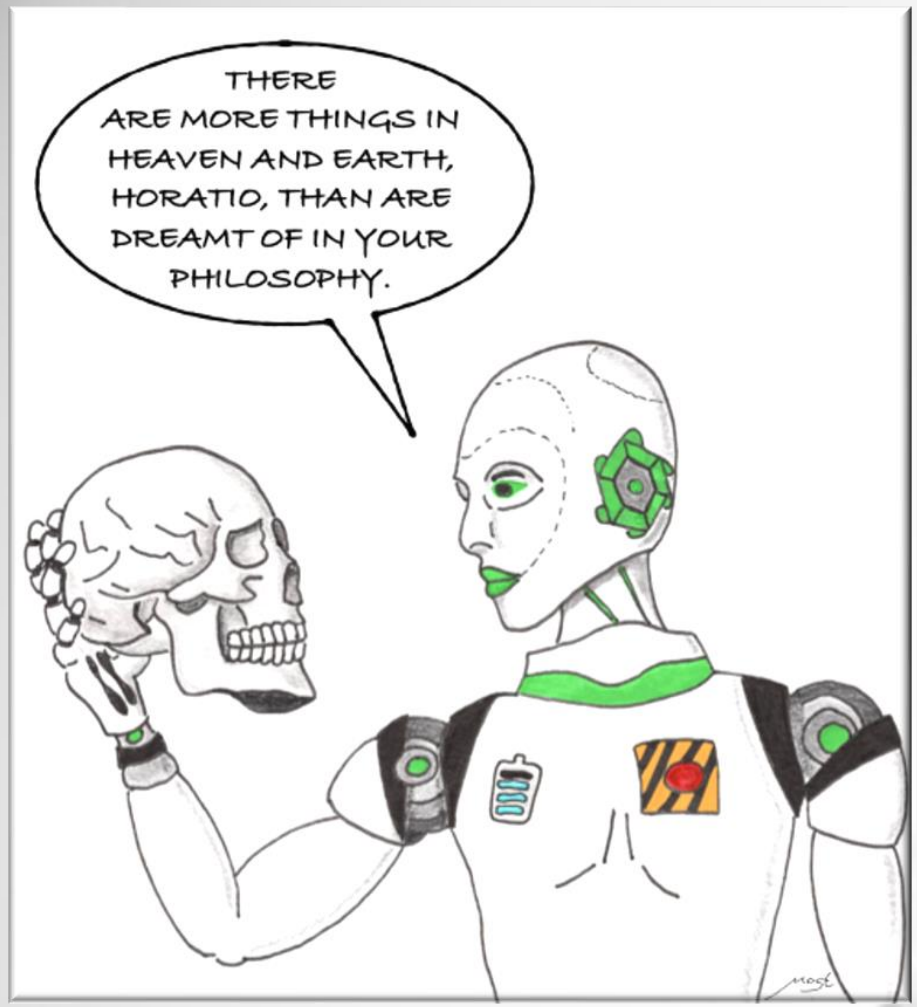


Artist: Moritz Strasser

Inbetweenism

How a philosophical framework may capture the varieties of social phenomena with GenAI

ANNA STRASSER (DenkWerkstatt Berlin, Germany)

Overview

MANY TERMS THAT PHILOSOPHERS PREVIOUSLY RESERVED FOR DESCRIBING THE DISTINGUISHING FEATURES OF HUMANS ARE NOW BEING APPLIED TO MACHINES, LEADING TO INTENSE DEBATES OVER SUCH NOTIONS AS UNDERSTANDING, KNOWLEDGE, REASONING, PHENOMENOLOGICAL CONSCIOUSNESS AND SOCIALITY.

I *How can we capture the variety of phenomena we are confronted with?*

II *How can we conceptualize INBETWEEN phenomena within a multidimensional spectrum?*

III *How to argue for a justifiable ascription practice?*

Slides can be downloaded at https://www.denkwerkstatt.berlin/ANNA-STRASSER/TALKS

Things don’t dichotomize

PHILOSOPHY POSES TOO DEMANDING CONDITIONS

Artist: Lorin Strasser

CLAIMS

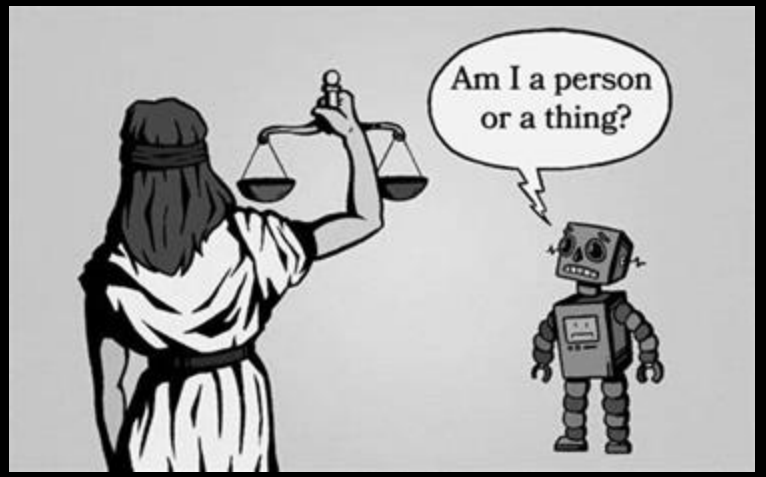
- ❖ philosophers tend to describe ideal cases that are rarely found in everyday life
- ❖ children, non-human animals, and robots (artificial agents) tend to fall through the conceptual net

➢ How can we expand or adopt the sophisticated terminology of philosophy to capture phenomena one finds in developmental psychology, animal cognition, and AI?

➢ GRADUAL APPROACHES & MINIMAL NOTIONS

(Strasser, 2006)

A conceptual problem



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

Is an LLM (or a robot developed with generative AI technology) a person or a thing?

- neither nor
- no philosophical terminology to describe what it is instead

WE CANNOT REDUCE ALL OF OUR INTERACTIONS WITH LLMS TO MERE TOOL USE

"It is neither quite right to say that all our interactions with artificial systems are mere tool use – nor is it quite right to say that these HMIs qualify as full-fledged social interactions. Neither ordinary concepts nor standard philosophical theorizing allow us to think well about these INBETWEEN phenomena."
(Strasser & Schwitzgebel, 2024, 197)

→ RETHINK OUR CONCEPTUAL FRAMEWORK
which so clearly distinguishes *between tools as inanimate things* and *humans as social, rational, and moral interaction partners*

Motivations

QUESTIONING THE DICHOTOMY BETWEEN ANIMATE AND INANIMATE

1

Western conception is just one conception of many

shintoism & animism

2

global rights-of-nature movement

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

- legal steps linking Western & Indigenous worldviews

3

notion of a social agent has proven to be changeable

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

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

What are we doing when we interact with LLMs?

DON'T ASK EITHER – OR QUESTIONS

Are interactions with LLMs social interactions?

NO
We are just playing with interesting tools.

YES
We act jointly with a social collaborator.

NEITHER NOR

TOOL USE

INBETWEEN PHENOMENA

FULL-FLEDGED SOCIAL INTERACTIONS

Two extreme positions

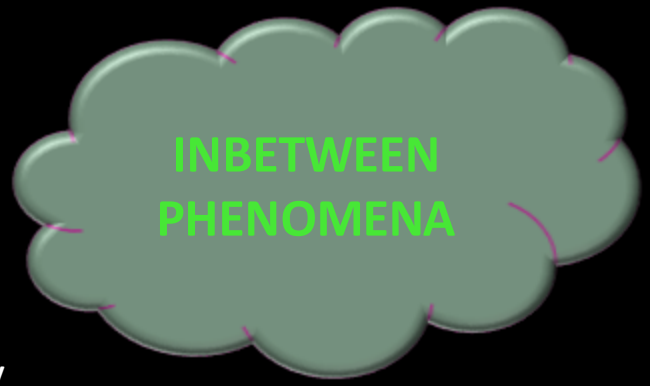
BASED ON THE DICHOTOMY BETWEEN ANIMATE & INANIMATE*

TOOL USE

Hard-core instrumentalists

- excluding the possibility that any artificial system could have a social status in an HMI

*INANIMATE → NOT SOCIAL



FULL-FLEDGED
SOCIAL INTERACTIONS

In-expectation of AGI view

- whole demanding package of conditions that we require from humans can, in principle, also be fulfilled by sophisticated machines

*ANIMATE → SOCIAL → ARTIFICIAL LIFE

PHILOSOPHY

→ restrictive use of concepts like agency, sociality, moral agency, moral patiency assumes that only living beings can qualify

When it comes to ethical questions

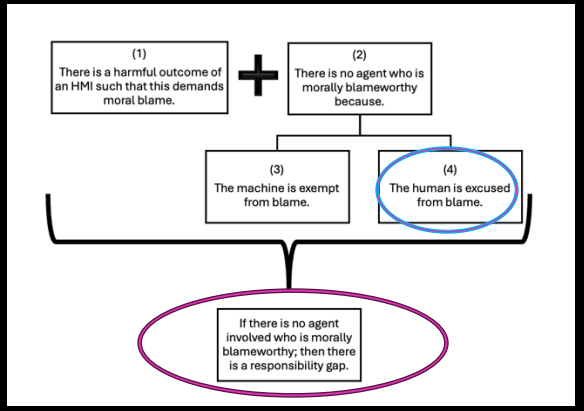
BOTH OPTIONS ARE NOT VERY ATTRACTIVE

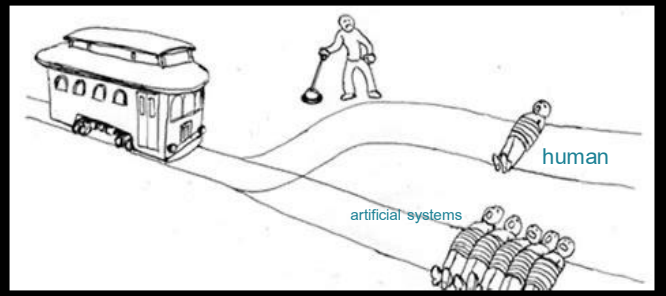
Hard-core instrumentalists

- either
  an increasing number of responsibility gaps
- or

revisions of established reasons for which humans can be excused from being responsible under certain circumstances in HMIs

- no straight-forward reasons to allow our interactions with artificial systems to be guided by moral or social norms





In-expectation of AGI view

- morally appropriate to sacrifice humans for machines
- risk of establishing a new rightless class of slaves
- need to revise our social practices of punishing

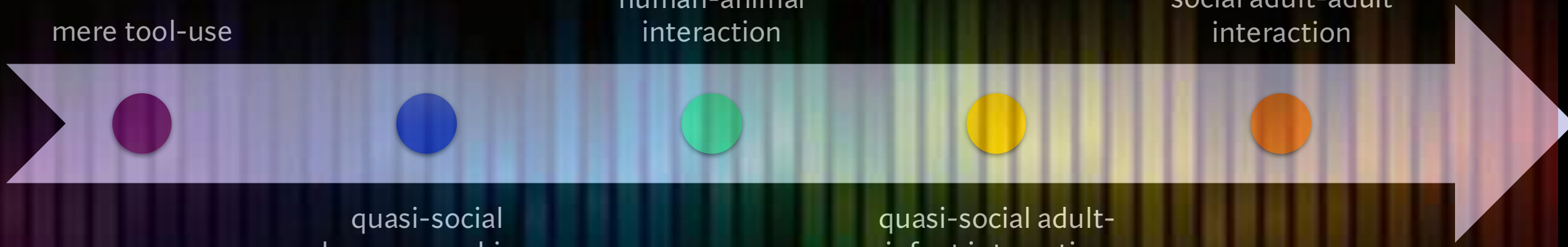
(Wilby & Strasser, 2024)

A multidimensional spectrum of social interactions

SINGLE-SIDED SOCIALITY

QUASI-SOCIAL ASYMMETRIC INTERACTIONS

FULL-BLOWN, INTELLECTUALLY DEMANDING, COOPERATIVE **SOCIAL INTERACTION**



Asymmetric distribution of abilities

PARADIGMATIC EXAMPLE OF SOCIAL INTERACTIONS THAT COULD BE APPLICABLE TO ARTIFICIAL SYSTEMS

NO NECESSITY OF AN EQUAL DISTRIBUTION OF ABILITIES AMONG ALL PARTICIPANTS

DEVELOPMENTAL PSYCHOLOGY

- social interactions action between adults and children
- children = socially interacting beings

ARTIFICIAL INTELLIGENCE

- human-machine interactions
- GenAI systems =?= socially interacting entities

ADULT & CHILD

ROBOT & HUMAN
LLM & HUMAN

DISTINCT TYPES OF ASYMMETRIC SOCIAL INTERACTIONS ARE CONCEIVABLE

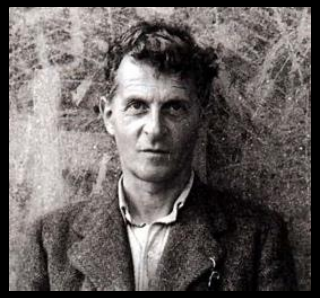
each type differs with respect to the proposed set of conditions

Towards a disjunctive conceptual framework

How can we capture the variety of phenomena we are confronted with?

Towards a disjunctive conceptual framework

How to characterize the many different instances in a multidimensional spectrum of social interactions



Wittgenstein, Ludwig. 2009.
Philosophical investigations.

INSTANCES STAND IN A RELATION OF FAMILY RESEMBLANCE

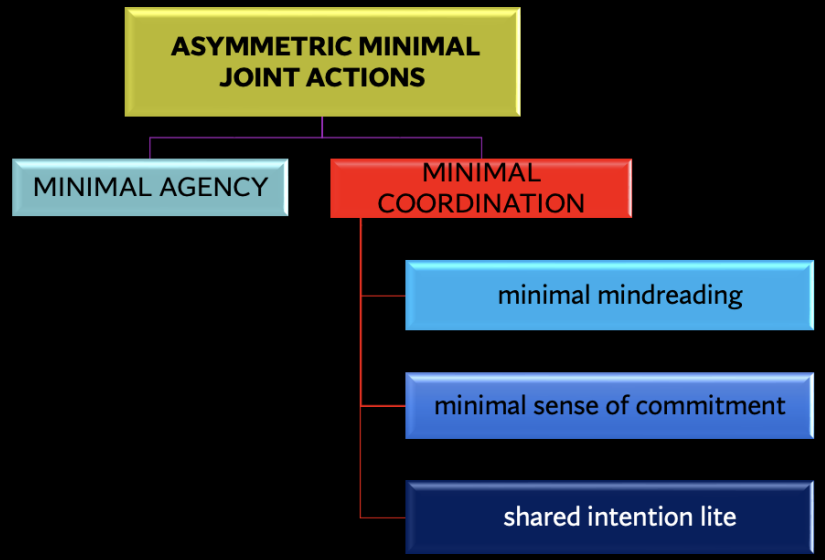
ALLOWING MULTIPLE REALIZATION

- a disjunctive conceptual framework does not require a whole package of conditions that necessarily co-occur, but allows for various combinations of conditions that can capture the diversity of phenomena

Towards a disjunctive conceptual framework

How to characterize the many different instances in a multi-dimensional spectrum of social interactions

MINIMAL APPROACHES



Stephen Butterfill & Ian Apperly (2013): minimal mindreading | John Michael et al. (2016): minimal sense of Commitment | Elisabeth Pacherie (2013): shared intention lite | Dominik Perler & Markus Wild (2022): simple mind

UTILIZING MINIMAL APPROACHES TO DESCRIBE VARIOUS SETS OF CONDITIONS

characteristic feature :

- questioning the necessity of some conditions
- allow for a less strong manifestation
- connect empirical findings and our common sense with theoretical work in philosophy

To characterize entailed conditions of artificial systems adequately, we need to enrich our terminology with further minimal notions.

Other disjunctive conceptual frameworks

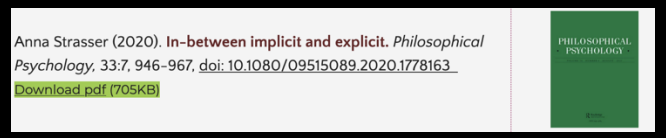
A FAMILIAR DISJUNCTIVE CONCEPTUAL FRAMEWORK CAN BE FOUND IN PSYCHIATRIC DIAGNOSTIC MANUALS

- both family resemblance & gradual variations play a role:
  - When diagnosed with a mental disorder, a person is assumed to have a certain number of symptoms, and it also matters how severe these symptoms are and how long the person is suffering from them …
  - ➢ Two persons can suffer from the same disorder even though they do not share the very same combination of symptoms.

➢ ARTIFICIAL SYSTEMS MAY QUALIFY AS QUASI-SOCIAL INTERACTION PARTNERS EVEN THOUGH THEY DO NOT FULFILL THE VERY SAME COMBINATION OF CONDITIONS AS HUMANS

Other disjunctive conceptual frameworks

DISJUNCTIVE CONCEPTUAL FRAMEWORKS ARE A GOOD TOOL TO CAPTURE THE VARIETIES OF PHENOMENA WE FIND IN EMPIRICAL RESEARCH



An either/or distinction between explicit and implicit processes comes with the consequence that not only different strengths of manifestations of conditions are neglected, but also interesting combinations of conditions are ignored. And for both we have empirical evidence.

- **questioning a dichotomous interpretation of two-system approaches by claiming that**

1. Properties vary by degrees

	system-one	**neglected INBETWEEN**	**system-two**
automatic	completely automatic	more-or-less automatic	non-automatic
controllable	no control	partial control	control
central accessibility	no central accessibility	limited central accessibility	central accessibility
access other information	informational encapsulated	limited accessibility	accessibility

2. Properties do not necessarily co-occur

- cognitive processes display a combination of properties:
  - conscious but uncontrollable, unintentional but still controllable, or efficient and intentional (Gawronski & Bodenhausen, 2011)
- automaticity is not necessary co-occurring with unconsciousness, unintentionality, efficiency, and uncontrollability

BY TAKING INTO ACCOUNT MANIFESTATIONS OF CONDITIONS IN VARIOUS STRENGTHS, LESS DEMANDING CONDITIONS CAN PROVE SUFFICIENT, AND BY QUESTIONING THE NECESSITY OF THE ENTIRE PACKAGE OF CONDITIONS, INTERESTING AND VARYING COMBINATIONS OF CONDITIONS CAN BE ACCOMMODATED.

Conceptualizing the multi-dimensional space of conceivable HMIs

A SPECTRUM RANGING FROM THE VERY FIRST WEAK INSTANCES OF QUASI-SOCIAL INTERACTIONS TO FULL-FLEDGED SOCIAL INTERACTIONS

very first weak instances of quasi-social interactions

- place relatively little demand on artificial interaction partners
- most minimal cases might not need
  - to have humanlike beliefs, desires, or self-generated goals
  - to be conscious
  - to understand much about their interaction partner
  - intend to communicate or cooperate

theoretically conceivable area

- no concrete hypothesis which of the many conceivable combinations of socio-cognitive abilities finally turn out to be sufficient
- advocating a gradual approach, the question of resemblance is a matter of degree

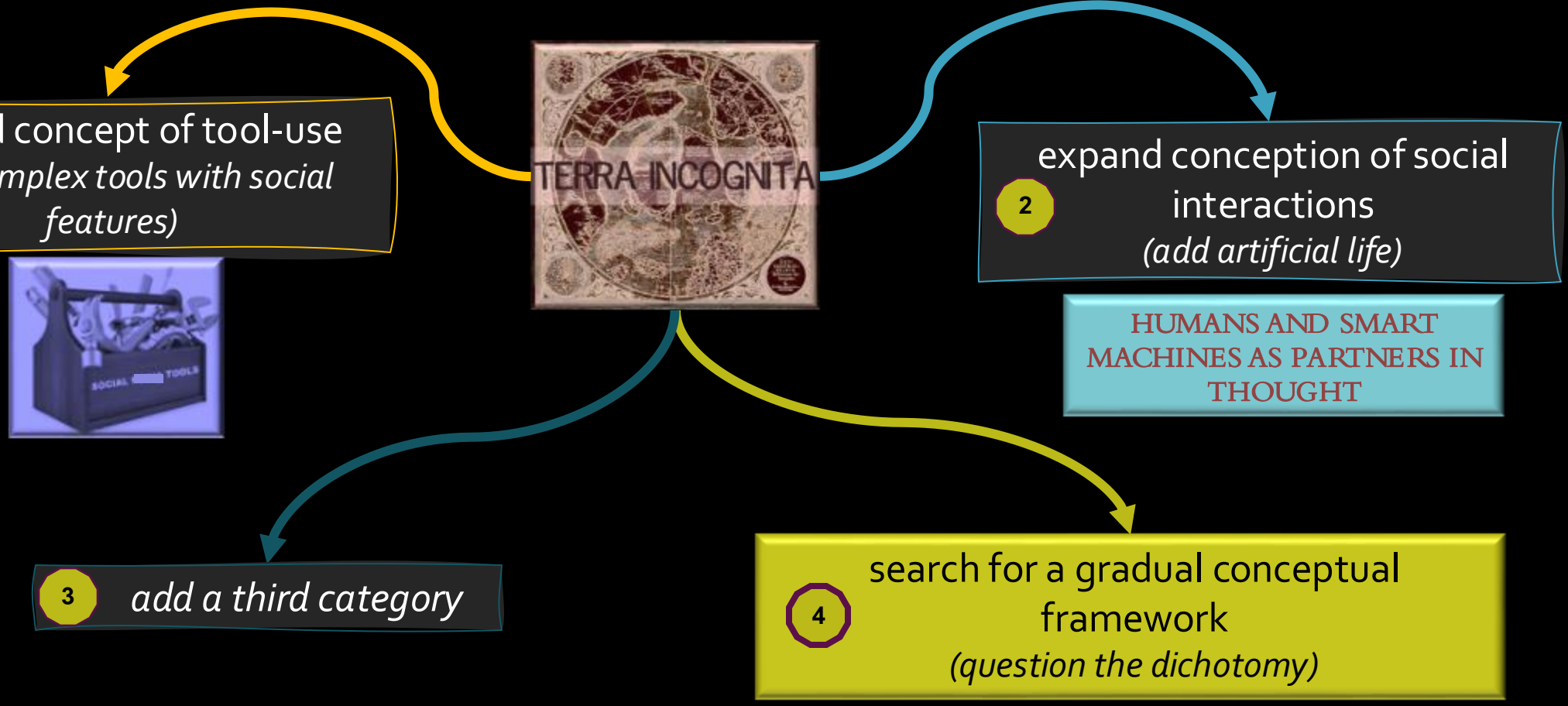
  - ➢ we cannot avoid a certain blurriness
  - ➢ be prepared for the possibility that there will be no clear-cut criteria to establish a sharp border

To qualify as quasi-social interaction partners, artificial systems must be structured to **not only** draw social behavior from their human partner **but also react to** that behavior in a way that solicits further social behavior and, importantly, these HMIs have to resemble social interactions as they transpire between two fully fledged social partners.

Other options?

CONCEPTIONS OF SOCIALITY ACCOUNT ONLY FOR LIVING BEINGS - NOT FOR ARTIFICIAL SYSTEMS

STATUS QUO: NO NOTIONS FOR IN-BETWEEN CASES



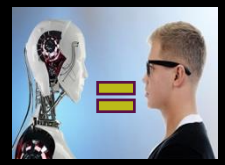
Objections

1 **emphasize the differences between humans & machines**

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

BUT

impossible to recognize potential multiple realizations of socio-cognitive capacities



2 **argue for similarities between humans & machines**

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

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.

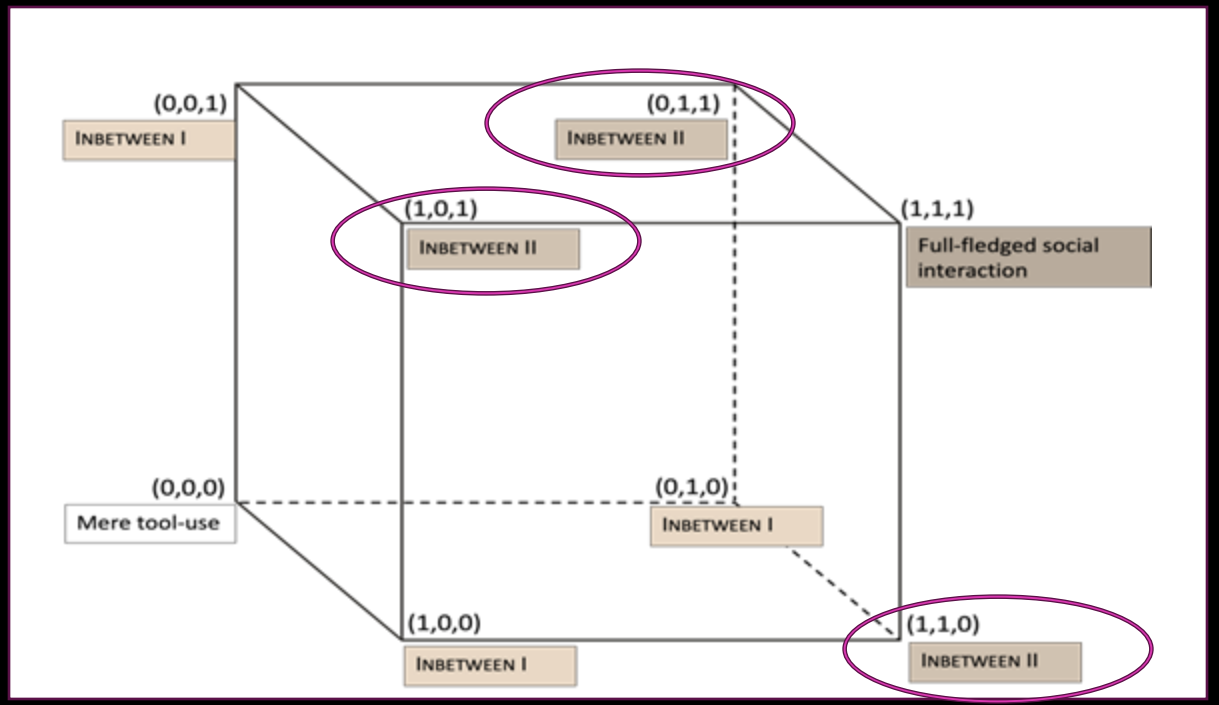
(Lemoine, 2022)

How to order conceivable instances in a multi-dimensional spectrum

EXCURSION INTO THE REALM OF COMBINATORICS

Seven instances of a simplified (inappropriate) disjunctive notion

		Condition 1	Condition 2	Condition 3
1	Full-fledged social interaction	1	1	1
2	INBETWEEN II	0	1	1
3		1	0	1
4		1	1	0
5	INBETWEEN I	0	0	1
6		0	1	0
7		1	0	0
	Tool use	0	0	0



We will have to be prepared for cases where we cannot answer the question of what types of quasi-social interaction partners are more social than other types of quasi-social interaction partners.

Multi-dimensionality is a complex matter

QUASI-SOCIALITY EXISTS ON A COMPLEX SPECTRUM

If we do not focus on adult humans as the only type of social partners

- we should expect that there are several dimensions along which we can characterize various instances of more or less social interactions

COMPLEX SOCIAL SKILLS WILL NOT EMERGE IN AN INSTANT

NOT

- *developmentally in humans*
- *phylogenetically in animal evolution*
- *technologically in the design of AI systems*

- Since social interchange is complex, there are multiple relevant dimensions of resemblance that concern the many presuppositions for agency and socio-cognitive abilities for sociality.

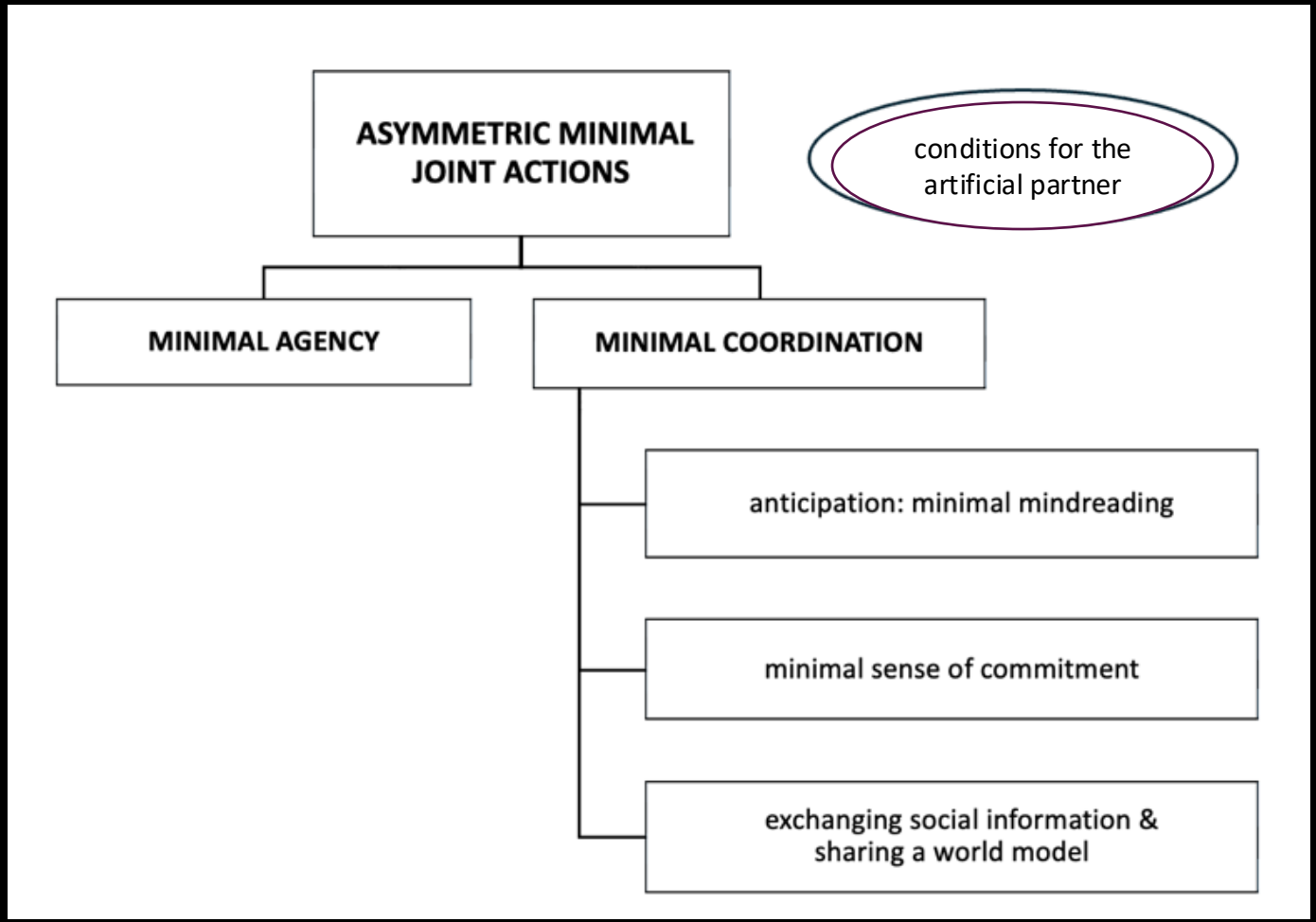
Asymmetric cases of joint actions

PARADIGMATIC EXAMPLE OF SOCIAL INTERACTIONS THAT COULD BE APPLICABLE TO ARTIFICIAL SYSTEMS

How to construct a minimal notion of an asymmetric joint action?

REQUIREMENTS FOR AGENCY & OTHER SOCIO-COGNITIVE ABILITIES

THAT CAN ENSURE THAT ARTIFICIAL AGENTS HAVE SUFFICIENT ABILITIES TO QUALIFY AS QUASI-SOCIAL INTERACTION PARTNERS



Justified ascriptions

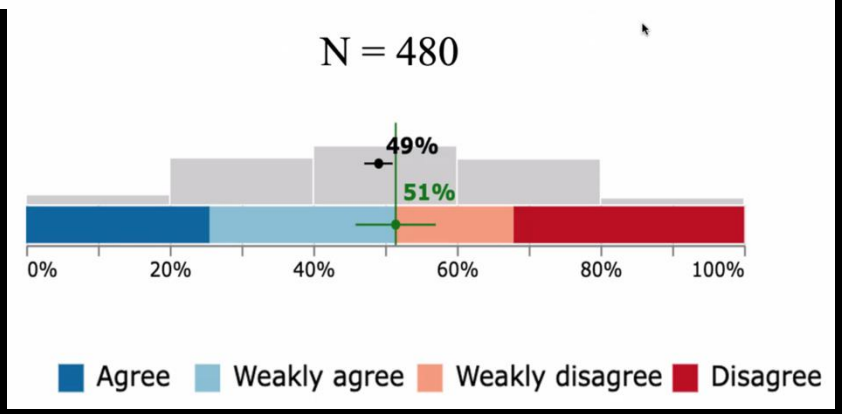
How to argue for a justifiable ascription practice?

WHAT DO NLP RESEARCHERS BELIEVE?
RESULTS OF THE NLP COMMUNITY METASURVEY

Julian Michael[1,2] Ari Holtzman[1] Alicia Parrish[4] Aaron Mueller[5] Alex Wang[3]
Angelica Chen[2] Divyam Madaan[3] Nikita Nangia[2]
Richard Yuanzhe Pang[3] Jason Phang[2] and
Samuel R. Bowman[2,3,4]

2022

Agree or disagree: Some generative models trained only on text, given enough data and computational resources, could understand natural language in some non-trivial sense.



(Michael et al., 2022)

Justified ascriptions

Neither the Turing test nor benchmarks deliver reliable reasons for socio-cognitive abilities

Turing test like measures

- test the ability of human evaluators to distinguish artificial systems from human interaction partners on the basis of their behavior

Benchmarks

- a machine that is able to solve presented tasks does not necessarily have to apply the supposed cognitive abilities to do so

RULE-FOLLOWING PARADOX

"This was our paradox: no course of action could be determined by a rule because any course of action can be made out to accord with the rule."

(Wittgenstein, 2003, § 201)

BENCHMARKS COME WITH CRITICAL ISSUES

- data contamination
- robustness of the results
- problems with flawed benchmarks

machine might make use of

- memorization
- shortcut learning
- subtle statistical associations

(Mitchell, 2023)

WE SHOULD BE CRITICAL OF WHETHER BENCHMARKS ACTUALLY MEASURE WHAT THEY CLAIM TO MEASURE

Beyond input-output patterns

WE NEED TO INVESTIGATE THE PROCESS BY WHICH THE PERFORMANCE IS ACHIEVED

MATHEMATICAL DESCRIPTIONS DO NOT YIELD USEFUL INSIGHTS INTO WHETHER THE PERFORMANCE IS DUE TO THE POSSESSION OF ANY SOCIO-COGNITIVE ABILITY

- no human-intelligible descriptions by which one could decide whether socio-cognitive abilities have emerged

detailed description of the human brain at the molecular and cellular levels

mathematical descriptions

of a huge composite function consisting of a complex sequence of linear and nonlinear transformations across many layers

Taking a physical stance towards human beings does not exclude the possibility that we are justified in taking an intentional stance towards them.

being able to give a mathematical description of neural nets does not yet exclude that they might possess socio-cognitive abilities

contra stating that because LLM's operations can be described by a mathematical description that refers to statistical calculations, linear algebra operations, or next-token predictions, those descriptions are also all we could ever ascribe to them

Interpretability techniques

INVESTIGATE THE INNER STRUCTURE OF NEURAL NETWORKS

- aim to uncover the causal mechanisms underlying LLMs' performance at a higher level

e.g. asking whether LLMs represent information, operate on representations, have activation patterns that realize abilities

PROBING, ATTRIBUTION, CAUSAL INTERVENTION

probing

- exploring what is encoded in a neural network
  - ➢ *statements about the likeliness of certain information to be represented in specific activation patterns*

attribution methods

- explore which parts of the input data a model relies on most for their outputs

causal intervention methods

- utilizing insights gained by probing & attribution to examine whether certain interventions can change the outputs of the system in a systematic way
  - ➢ determine the causal role played by a representation in the processing of a system

A very accessible presentation of the details of such approaches can be found in *A Philosophical Introduction to Language Models* (Millière & Buckner, 2024b, 2024a)

TWO DIFFICULTIES

- techniques are mostly practiced with toy models
  - ➢ wait until they are applied to large language models
- rely on operationalizable theories of all the abilities we want to ascribe to LLMs

Plea for cross-disciplinary approaches

- ascription of properties and socio-cognitive abilities to artificial systems cannot be clarified by computer science alone

- purely philosophical theorizing also has not yet led to a practical strategy of how one can justifiably argue for certain ascriptions.

At this point, one could despair and say that we are staring into an abyss and that there is little hope that we will ever be able to build conceptual bridges in the foreseeable future that will allow us to ascribe certain properties and abilities to artificial systems clearly.

This uncertainty regarding the justified attribution of properties and capabilities motivates an urgent need for cross-disciplinary cooperation which might have the potential to suggest a commonly agreed-on practice of how one can adequately describe the status of artificial systems in HMIs.

Conclusion

I

How can we capture the variety of phenomena we are confronted with?

→ assume a multi-dimensional spectrum that includes the inbetween phenomena that we cannot describe with our standard terminology

II

How can we conceptualize INBETWEEN phenomena within a multidimensional spectrum?

→ establish a disjunctive conceptual framework that entails new minimal notions

III

How to argue for a justifiable ascription practice?

→ challenging endeavour that cannot be met by computer science or by philosophy alone

→ plea for a cross-disciplinary approach

Slides can be downloaded at https://www.denkwerkstatt.berlin/ANNA-STRASSER/TALKS

Acknowledgment

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

Daniel Dennett

Eric Schwitzgebel

Joshua Rust

Steven Butterfill

Mike Wilby

DigiDan

Thank you !

References

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

Gawronski, B., & Bodenhausen, G. V. (2011). The Associative–Propositional Evaluation Model. In *Advances in Experimental Social Psychology* (Vol. 44, pp. 59–127). Elsevier. https://doi.org/10.1016/B978-0-12-385522-0.00002-0

Gunkel, D. J. (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

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

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

Michael, J., Holtzman, A., Parrish, A., Mueller, A., Wang, A., Chen, A., Madaan, D., Nangia, N., Pang, R. Y., Phang, J., & Bowman, S. R. (2022). *What Do NLP Researchers Believe? Results of the NLP Community Metasurvey* (No. arXiv:2208.12852). arXiv. https://doi.org/10.48550/arXiv.2208.12852

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

Millière, R., & Buckner, C. (2024a). *A Philosophical Introduction to Language Models -- Part I: Continuity With Classic Debates* (No. arXiv:2401.03910). arXiv. https://doi.org/10.48550/arXiv.2401.03910

Millière, R., & Buckner, C. (2024b). *A Philosophical Introduction to Language Models - Part II: The Way Forward* (No. arXiv:2405.03207). arXiv. https://doi.org/10.48550/arXiv.2405.03207

Mitchell, M. (2023). AI's challenge of understanding the world. *Science*, *382*(6671), eadm8175. https://doi.org/10.1126/science.adm8175

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

Perler, D., & Wild, M. (Eds.). (2022). *Der Geist der Tiere: Philosophische Texte zu einer aktuellen Diskussion* (6. Auflage). Suhrkamp.

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

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

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

Strasser, A. (2020). In-between implicit and explicit. *Philosophical Psychology*, *33*(7), 946–967. https://doi.org/10.1080/09515089.2020.1778163

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

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.

Wittgenstein, L. (2003). *Philosophische Untersuchungen* (J. Schulte, Ed.). Suhrkamp.