

TUESDAY, AUGUST 25TH

17:30-19:25 Session B6: Philosophy of Mind and Action: Actions

LOCATION: [Multiusos Building - Room 1210](#)

18:10 [Anna Strasser](#)

Towards a Multidimensional Spectrum of Agency

ABSTRACT. Characterizing the role generative AI systems can play in interactions with humans, we face controversial positions. Disputes revolve around whether AI systems can qualify as social interaction partners or should rather be described as tools. Focusing on agency, particularly the ability to act jointly, this talk will explore what forms of agency are conceivable within a multidimensional spectrum. With reference to the manifold forms of agency, ranging from animate, sentient, and volitional agency of self-moving creatures to forms of inanimate agency, one can find in philosophy, I will argue that in asymmetric joint actions, we do not need to pose the very same conditions on all involved participants and sketch how AI systems may qualify as interaction partners with a minimal form of agency and socio-cognitive abilities.

Towards a multidimensional spectrum of agency

Anna Strasser
DenkWerkstatt Berlin

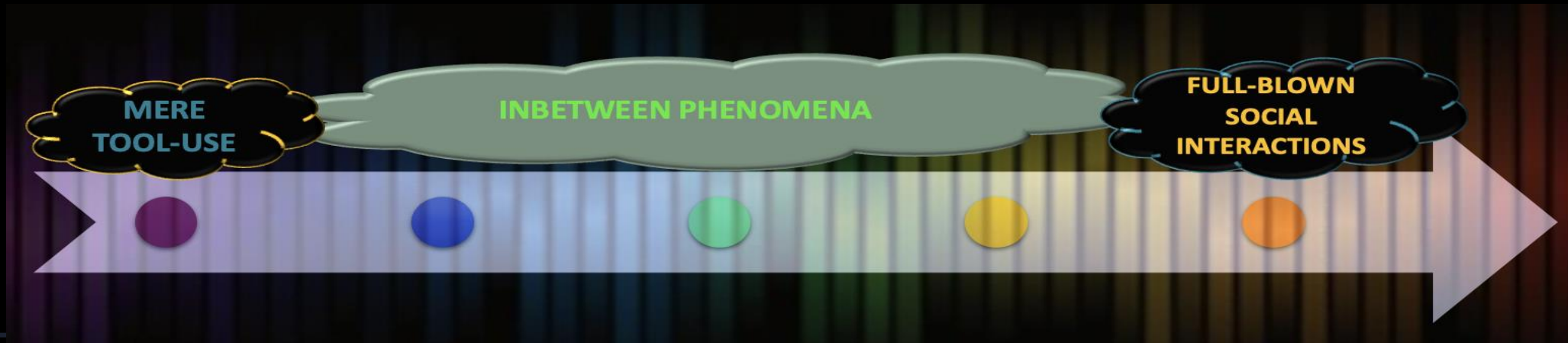
Intro

In light of debates about generative AI systems, we face a range of controversial viewpoints on how to characterize the role of these novel entities.

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

- agency, intelligence, consciousness / beliefs, mental representations, systematic generalization abilities, understanding

The disputes here often revolve around whether such human-machine interactions should be described as tool use or as social interactions.



Overview

Controversial debates on necessary conditions and extension of agency

Ascribing social agency to generative AI systems?

Future Research

MULTIDIMENSIONAL SPECTRUM OF AGENCY
WITH
MANIFOLD FORMS OF ANIMATE, SENTIENT,
AND VOLITIONAL AGENCY OF SELF-MOVING
CREATURES, AS WELL AS FORMS OF
INANIMATE AGENCY (Hacker, 2007)

JOINT AGENCY MAY POSE LESS-
DEMANDING CONDITIONS ON ONE
INTERACTION PARTNER IN
ASYMMETRIC CASES OF
INTERACTIONS BETWEEN UNEQUAL
PARTNERS

CONCERNING POTENTIAL LIMITATIONS
OF AI SYSTEMS & ABOUT WHAT
INFERENCES ONE MAY DRAW FROM
CERTAIN INSTANCES

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



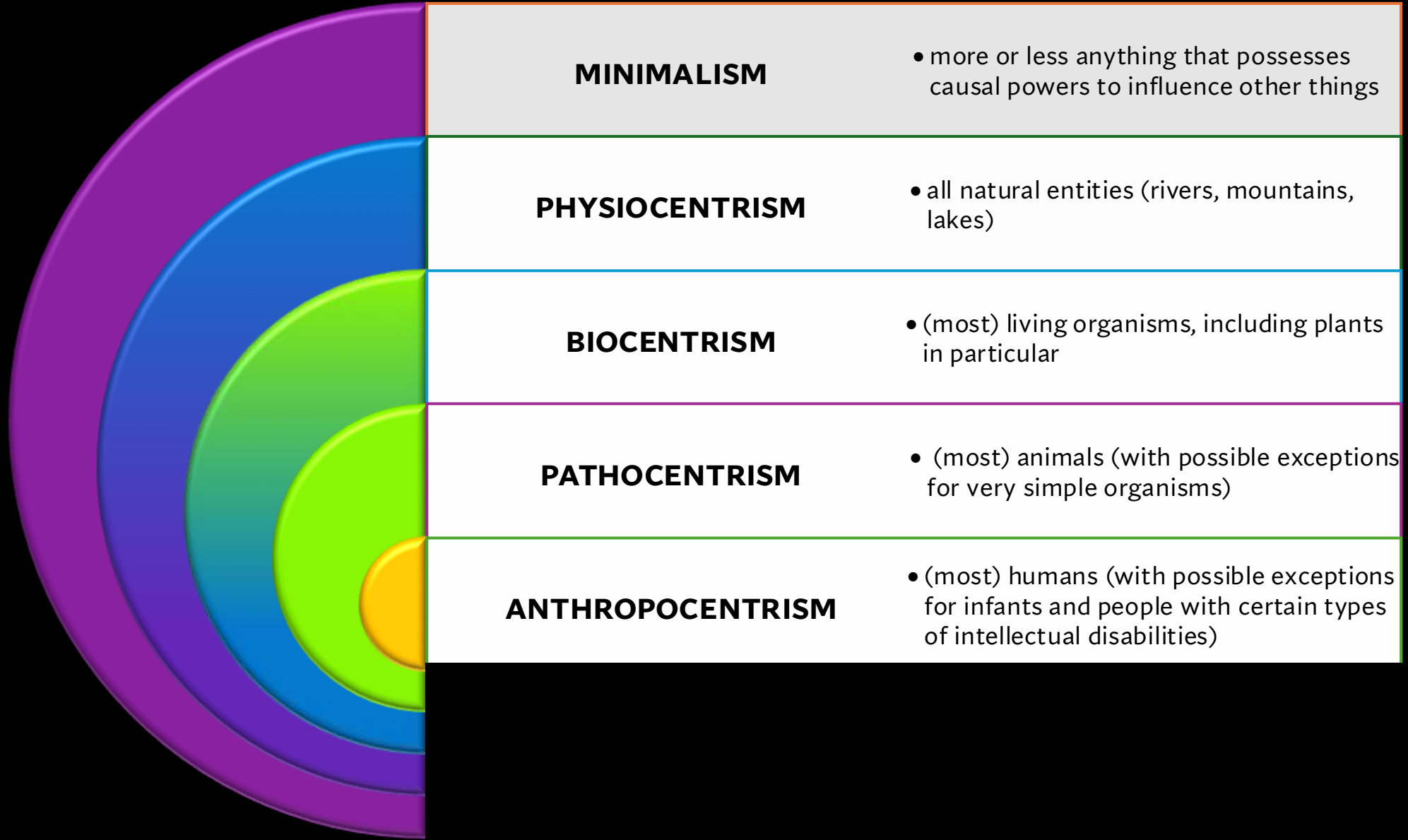
Agency

WHAT FORMS OF AGENCY ARE CONCEIVABLE IN A MULTIDIMENSIONAL SPECTRUM OF AGENCY?

manifold forms of animate, sentient, and volitional agency of self-moving creatures,
as well as forms of inanimate agency (Hacker, 2007)



Extension of agency



Consciousness

sentence & vulnerability

- embodiment plays an essential role in 'genuine' agency (Varela et al., 2017)
- Are only living meat machines conscious? (Block, 2025)

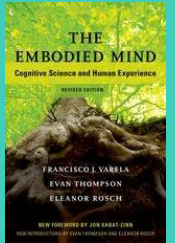
CellPress

Trends in
Cognitive Sciences

Opinion

Can only meat machines be conscious?

Ned Block 1*



embodiment, autonomy,
personhood, aliveness, and
responsibility

Intelligence

characterizing learning abilities, forms of
rationality (instrumental & reflective rationality)

- **minimal:** cognitive-conative abilities to grasp relevant information & flexibly adapt their behavior to changes of the environment to pursue goals (instrumental rationality (Schmidtz, 1994))
- **more demanding:** make predictions & plans, decide, choose, evaluate, be guided by reasons (instrumental & reflective rationality (Schmidtz, 1994))
- **most demanding:** recognize moral reasons
- **+ controversial theories regarding mental states:** propositional states, beliefs, language, representations (of world models & goals)

Goal-orientation

with or without intentions, desires, and the ability to generate
goals

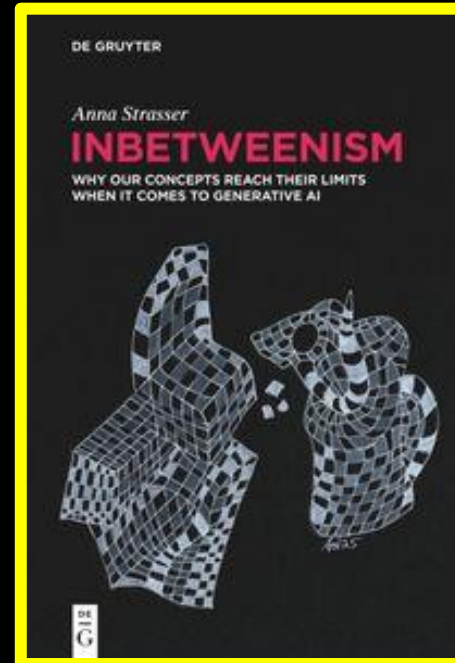
ABILITY TO BRING ABOUT CHANGES

What role could or should we ascribe to GenAI systems?

CAN AI SYSTEMS SERVE AS SOCIAL COUNTERPARTS?

Hardcore instrumentalist view

no agency
no intelligence
no consciousness



- AI systems cannot have a social status in human-machine interactions.

UNDERATTRIBUTION

- challenged by observations that these systems are significantly infiltrating our social spheres

In expectation of AGI view

full-fledged
agency
intelligence
consciousness



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

OVERATTRIBUTING

- seem to ignore differences between machines & humans
 - possibility of artificial life

No hope for a consensus?

CONCEPTUAL PROBLEM:

there are now phenomena in our world for which we lack adequate notions

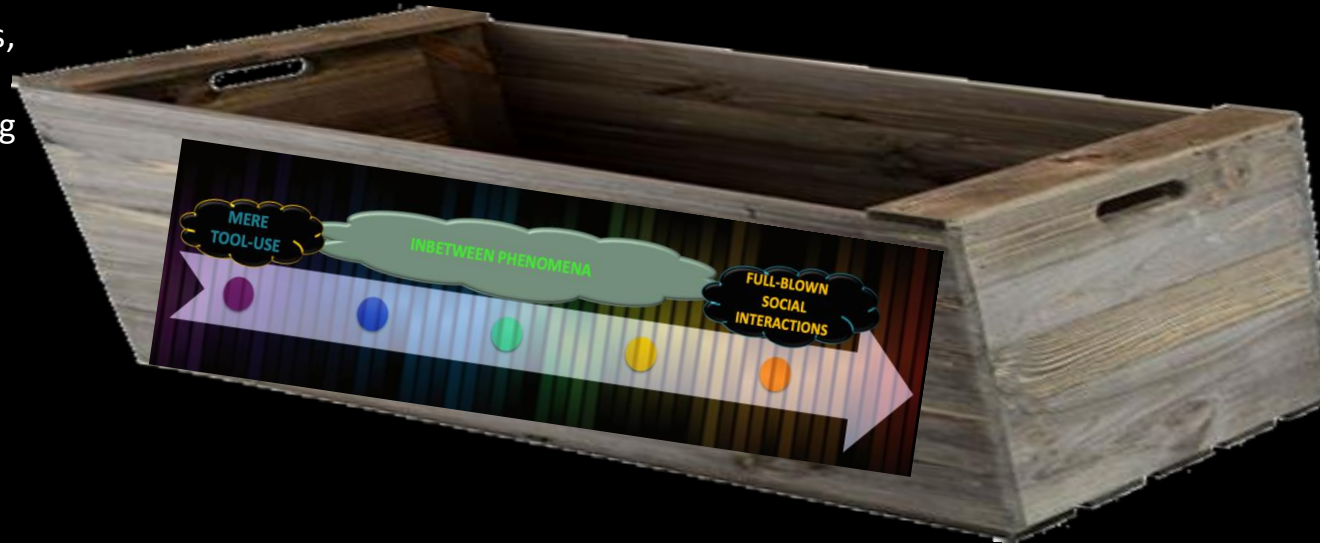
Neither standard philosophical theorizing nor dichotomous ordinary concepts enable us to think well about these in-between phenomena.



people fight to rescue notions, which in their view should remain reserved for describing distinguishing features of humans



social reality in which people entertain all kinds of social interactions with GenAI Systems



QUASI-SOCIAL ASYMMETRIC SOCIAL INTERACTIONS

mere tool-use

quasi-social
human-animal
interactionsocial adult-adult
interactionSINGLE-SIDED
SOCIALITYquasi-social
human-machine
interactionquasi-social adult-
infant interactionFULL-BLOWN SOCIAL
INTERACTION

ASCRIBING A SOCIAL ROLE TO ARTIFICIAL COUNTERPARTS, WE NEED TO CONSIDER BOTH THE SOCIAL REALITY AND THE PROPERTIES OUR INTERACTION PARTNERS HAVE.

Multidimensionality is a complex matter

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*

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

**JOINT ACTION FIRST & ASYMMETRICAL DISTRIBUTION OF
CONDITIONS MAKE ROOM FOR INSTANCES THAT DESCRIBE
CERTAIN HUMAN-MACHINE INTERACTIONS**

Asymmetric cases of joint actions

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

INTELLECTUALIST CONCEPTIONS

- ❖ philosophers tend to describe ideal cases



FULL-FLEDGED AGENCY BY
DONALD DAVIDSON

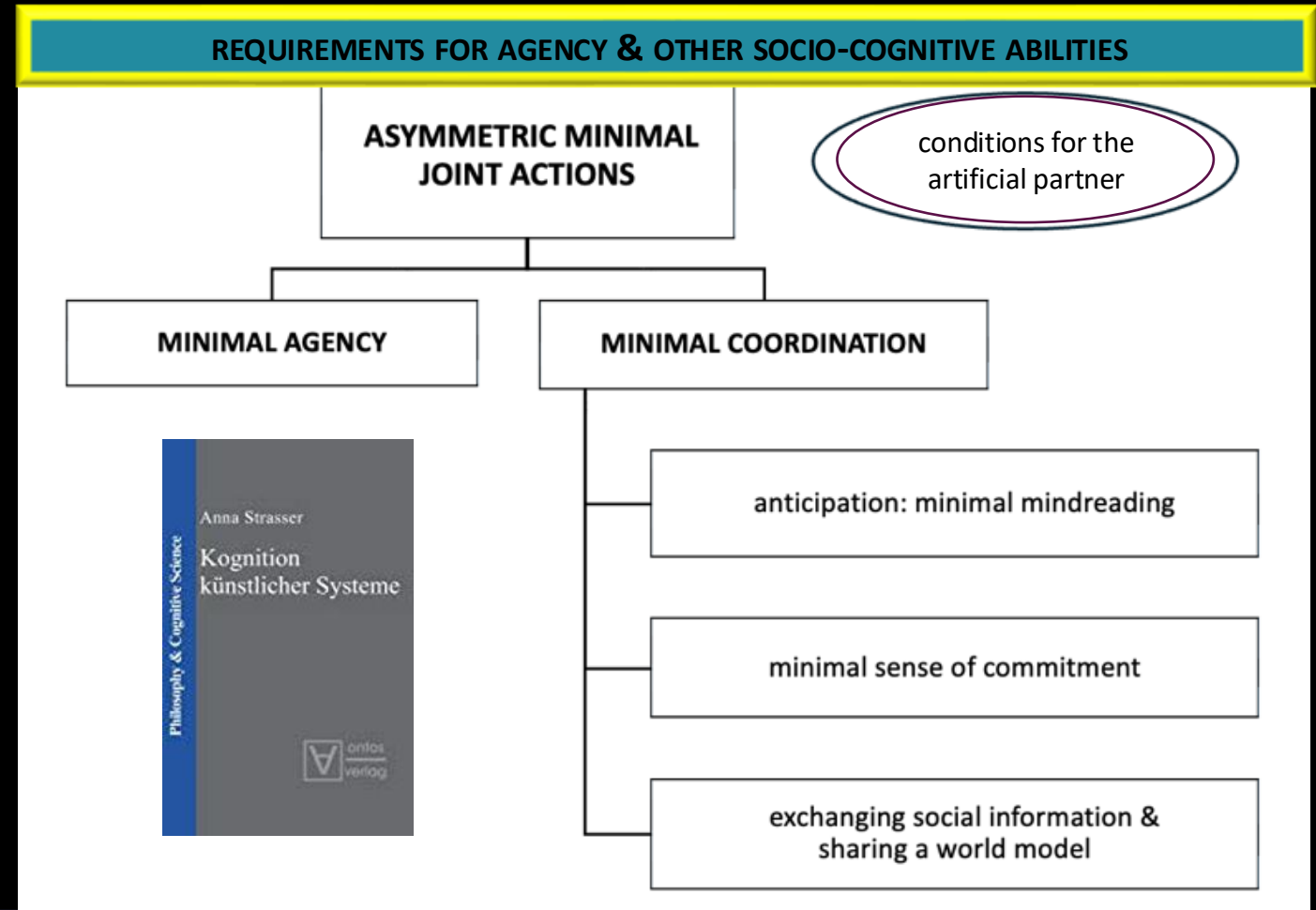
- ❖ children, non-human animals, and artificial systems tend to fall through the conceptual net



Artist: Lorin Strasser



GRADUALIST APPROACHES & MINIMAL NOTIONS



Does this help regarding the ascription problem?



WE DO NOT KNOW WITH CERTAINTY WHETHER THE CURRENT AI SYSTEMS FULFILL THE PROPOSED CONDITIONS

Any application of a conceptual framework raises empirical questions, namely, questions about the extent to which certain conditions (minimal abilities and properties) are fulfilled by the entity in question.

I put some hope on the further progress of so-called interpretability technics, and expect that further analysis of our social reality will add reasons to see that there are HMI's that can best be described as asymmetric joint actions.

I can avoid the critical (and, from my view, hopeless) debate concerning the ascription of consciousness.

MY CLAIM

There are forms of agency and intelligence that do not require consciousness —
the three features are not always served in one package!

(Strasser 2026b)

FUTURE RESEARCH

- establish cross-disciplinary approaches considering social and technical factors of our ascription practice concerning the social status of artificial systems

→ may lead to a lucky marriage of ideas of the property-based and relational approaches

Future research

HMIs can qualify as instances in this multidimensional spectrum of social interactions as asymmetric joint actions that happen in the domain of working

POTENTIAL DISRUPTIVE EFFECTS
OF GENAI IN VARIOUS SOCIAL
DOMAINS.

SOME OF OUR SOCIAL INTERACTIONS

- may require being less asymmetric,
- may demand that both interaction partners exhibit emotions and reflective rationality

Claims like

AI THERAPISTS CANNOT
CONTRIBUTE TO A THERAPEUTIC
RELATIONSHIP IN WHICH
EMOTIONAL CO-REGULATION IS
HAPPENING

AI FRIENDS CONSTITUTE A NEW FORM
OF PARASOCIAL FRIENDSHIP, AN
INTERESTING MIXTURE OF UTILITY &
PLEASURE FRIENDSHIP THAT HAS THE
POTENTIAL TO BE RATHER TOXIC.

Future research

My suggestion to imagine a multidimensional spectrum of agency provoked all kinds of objections. Many people accuse me of watering down the concept of agency.

Maybe there's something to that, but my intention is not to blur the distinctions between concepts.

SPEAKING OF DIFFERENT INSTANCES OF AGENCY, WE CAN ALSO MORE CLEARLY DISTINGUISH WHICH INFERENCES SHOULD BE DRAWN FROM WHICH INSTANCE.

- we can hold the agents responsible for their actions
 - implies moral agency and moral patiency

**FULL-FLEDGED
INDIVIDUAL
AGENCY**

**FULL-BLOWN
SOCIAL
INTERACTIONS**

- both interactants are morally responsible for the outcome of the interactions
- social & moral norms apply

**INSTANCES OF
ASYMMETRIC
INTERACTIONS**

**MINIMAL
SOCIAL
AGENCY**

- argue for a somehow social status AI systems may have in interactions with humans

- no moral agency
- but can imply moral patiency (→ non-human animals)

Conclusion

Since there are controversial debates on necessary conditions and extension of agency, I suggested to imagine a multidimensional spectrum of agency. I aim to work on a map that shows how all the different concepts of agency, I can get hold of, can be characterized as instances in this spectrum.

➤ **If you are in favor of a certain conception of agency, I am happy to receive an email with some references.**

I delivered a sketch of how one could describe certain interactions between humans and artificial systems as asymmetric joint actions, claiming that apparent necessary conditions of full-blown social interactions can be questioned for asymmetric interactions with unequal partners.

➤ Which led me to list minimal necessary conditions for artificial quasi-social interaction partners. This list frees AI systems not only from being able to set goals as long as they are able to contribute to the joint action but also from being conscious and possessing reflective rationality.

Since this is still an ongoing project, I concluded with two ideas for future research:

1. I highlighted social interactions that could illustrate the limitations of current AI systems, such as AI therapists and companions.
2. I believe it is insightful to examine the inferences that arise from specific instances within this multidimensional spectrum—inferences that may also shed light on the underlying motivations for why a concept of action was developed for that instance.

- Agüera y Arcas, B. (2022). Do Large Language Models Understand Us? *Daedalus*, 151(2), 183–197. https://doi.org/10.1162/daed_a_01909
- 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>
- Block, N. (2025). Can only meat machines be conscious? *Trends in Cognitive Sciences*, S1364661325002347. <https://doi.org/10.1016/j.tics.2025.08.009>
- Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness* (arXiv:2308.08708). arXiv. <https://doi.org/10.48550/arXiv.2308.08708>
- Davidson, D. (1980). *Essays on Actions and Events*. Oxford University Press. <https://doi.org/10.1093/0199246270.001.0001>
- Hacker, P. M. S. (2007). *Human Nature: The Categorical Framework* (1st ed.). Wiley. <https://doi.org/10.1002/9780470692165>
- Harding, J. (2023). Operationalising Representation in Natural Language Processing. *The British Journal for the Philosophy of Science*, 728685. <https://doi.org/10.1086/728685>
- Hashim, S. (2024, December 5). *Anthropic has hired an “AI welfare” researcher*. <https://www.transformernews.ai/p/anthropic-ai-welfare-researcher>
- Herrmann, D. A., & Levinstein, B. A. (2025). Standards for Belief Representations in LLMs. *Minds and Machines*, 35(1), 5. <https://doi.org/10.1007/s11023-024-09709-6>
- Kohn, E. (2013). *How forests think: Toward an anthropology beyond the human*. University of California Press.
- Lake, B. M., & Baroni, M. (2023). Human-like Systematic Generalization through a Meta-learning Neural Network. *Nature*, 623, 115–121. <https://doi.org/10.1038/s41586-023-06668-3>
- Latour, B. (2005). *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press Oxford. <https://doi.org/10.1093/oso/9780199256044.001.0001>
- Lemoine, B. (2022, June 11). Is LaMDA Sentient? — An Interview. *Medium*. <https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917>

- Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). *Taking AI Welfare Seriously* (arXiv:2411.00986). arXiv. <https://doi.org/10.48550/arXiv.2411.00986>
 - Open-AI. (n.d.). *Planning for AGI and beyond*. Retrieved May 6, 2024, from <https://openai.com/index/planning-for-agi-and-beyond>
 - Schmidt, D. (1994). Choosing Ends. *Ethics*, 104(2), 226–251. <https://doi.org/10.1086/293599>
 - Strasser, A. (2006). *Kognition künstlicher Systeme*. De Gruyter. <https://doi.org/10.1515/9783110321104>
 - Strasser, A. (2026a). *Inbetweenism: Why our concepts reach their limits when it comes to Generative AI*. De Gruyter. <https://doi.org/10.1515/9783111438078>
 - Strasser, A. (2026b). What Smart AIs Can Do without Consciousness. *Journal of Artificial Intelligence and Consciousness*, 13(01), 37–62. <https://doi.org/10.1142/S2705078526400023>
 - Strasser, A., & Schwitzgebel, E. (2024). Quasi-sociality: Toward Asymmetric Joint Actions. In *Anna's AI Anthology. How to live with smart machines?* (pp. 195–211). xenomoi Verlag.
 - Varela, F. J., Thompson, E., & Rosch, E. (2017). *The Embodied Mind: Cognitive Science and Human Experience*. The MIT Press. <https://doi.org/10.7551/mitpress/9780262529365.001.0001>
 - Vesper, C., Butterfill, S., Knoblich, G., & Sebanz, N. (2010). A minimal architecture for joint action. *Neural Networks, Social Cognition: From Babies to Robots*, 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*. *Intelligencer*. <https://nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html>
-

What should count as a justified ascription practice?

PROPERTY-BASED VERSUS RELATIONAL APPROACHES

PROBLEMS OF PROPERTY-BASED APPROACHES

- no agreed-on strategy for deciding whether an entity really possesses a required property
 - we haven't solved the Other Mind problem
 - nor have we developed agreed-on measurement practices
 - only looking at outputs does not inform us of whether the processing that leads to certain outputs captures conditions we pose on thinking, reasoning, or other mental processes
 - huge variety of positions that differ in what specific properties are required for abilities and properties and how they could be operationalized

PROPERTIES WHICH ARE SUPPOSED TO ESTABLISH MORAL OR SOCIAL STATUS ARE BOTH METAPHYSICALLY AND EPISTEMOLOGICALLY PROBLEMATIC

PROBLEMS OF RELATIONAL APPROACHES

- taking our attitude towards an interaction partner as a reason for ascribing a social role to them might in some cases be inappropriate,
 - lead to the risk of being overinclusive and arbitrary

IT IS HARD TO ANTICIPATE AND DETERMINE WHAT BELONGS TO OUR SOCIAL REALITY