slides can be downloaded at

https://www.denkwerkstatt.berlin/ANNA-STRASSER/TALKS

How to argue for agency of artificial systems

– joint actions first –

Anna Strasser

DenkWerkstatt Berlin / LMU Munich

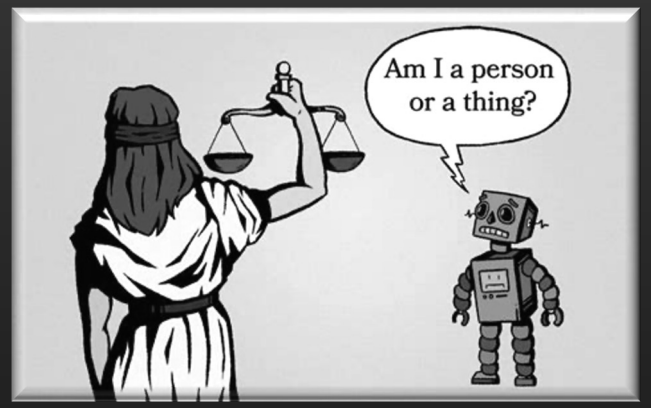
THE 11TH EUROPEAN CONGRESS OF ANALYTIC PHILOSOPHY

EUROPEAN SOCIETY FOR ANALYTIC PHILOSOPHY

Can smart machines act?

Can all types of human-machine interactions be reduced to simple tool use?

Are all artificial systems mere inanimate things that can
only behave as they were designed to
but cannot act themselves?

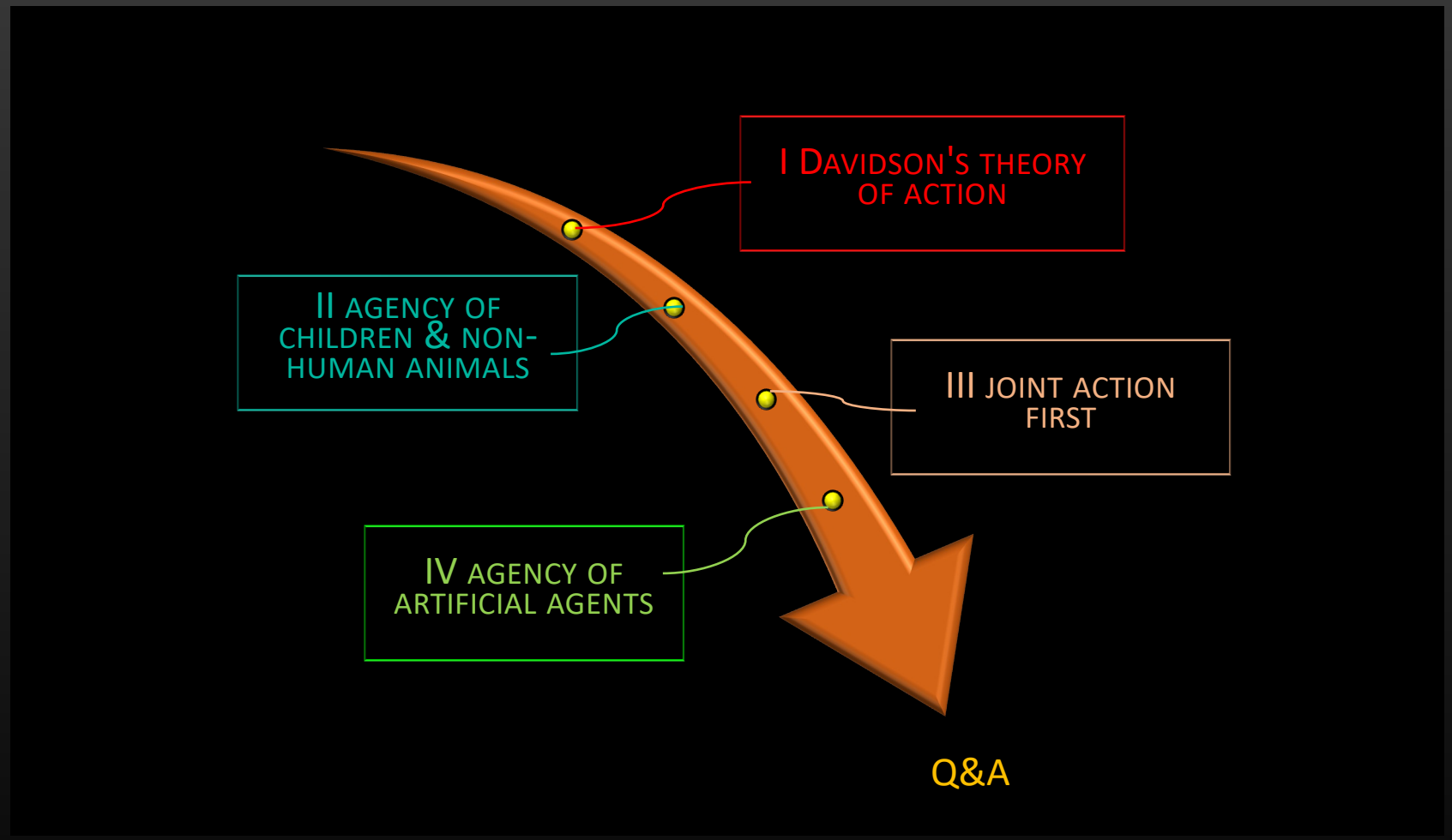


NOT ONLY SOPHISTICATED ADULT HUMANS
BUT ALSO CHILDREN, NON-HUMAN ANIMALS, AND ARTIFICIAL SYSTEMS ARE ABLE TO ACT

extended account of action that can consider artificial systems as acting participants in joint actions

DENKWERKSTATT BERLIN

Overview



20-25 min talk + 10 -15 Q&A

DENKWERKSTATT BERLIN

1. Individual agency á la Davidson

THE NECESSITY OF A COMPLEX SUITE OF CONCEPTUAL RESOURCES

constitutive relations holding between
- propositional attitudes & their contents
- language
- intentional agency
- interpretation

} sharply separate off 'the beasts' from rational animals such as humans

FULL-BLOWN INTENTIONAL AGENCY
requires intentional action
to be carried out by an entity **with**
an integrated, holistic set of
propositional attitudes

"The intrinsically holistic character of the propositional attitudes makes the distinction between having any and having none dramatic!"

Donald Davidson (1963, 1971, 1980, 1982, 1984, 2001)

Intellectualist conceptions of intentional action

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

too demanding conditions whose necessity can be questioned

Strasser & Wilby (2023)

Objections

Empirical-based

DEVELOPMENTAL & COMPARATIVE PSYCHOLOGY

counterexamples

MULTIPLE REALIZATIONS OF AGENCY IN INFANTS & NON-HUMAN ANIMALS

Perler & Wild 2005, Premack & Woodruff 1978, Heyes 2014/2015, Vesper et al. 2010

→ NOT ONLY CONCEPTUALLY SOPHISTICATED HUMANS CAN ACT

Conceptual-based

ONTOGENETICS & PHYLOGENETICS

counterexamples

SHIFT FROM NON-INTENTIONAL TO INTENTIONAL IS **GRADUAL & PARTLY LEARNABLE**

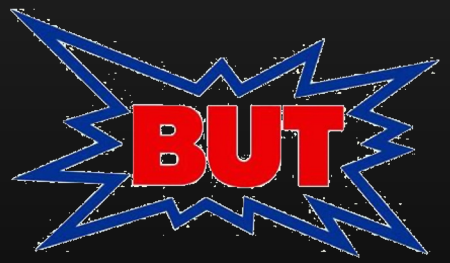
- Ontogenetic case: Perner 1991, Tomasello 2008
- Phylogenetic case: Sterelny 2014, Henrich 2016

→ DAVIDSONIAN 'ALL-OR-NOTHING' DRAMATIC DIVIDE IS IMPLAUSIBLE

DENKWERKSTATT BERLIN

Biological conceptions of intentional agency

IF
any kind of agency requires consciousness including internal states
(emotional, mental & conscious states)
THEN
a biological make-up appears as a necessary condition



IF ARTIFICIAL SYSTEMS CAN ONLY BEHAVE NOT ACT

BECAUSE THEY LACK THE BIOLOGICAL MAKE-UP THEN

→ **EVERY HUMAN-MACHINE INTERACTION SHOULD BE UNDERSTOOD AS MERE TOOL-USE** ←

DENKWERKSTATT BERLIN

Multiple realizations

Why should we disqualify machines because they are not living, biological beings?

What about assuming, that the way living beings fulfill the conditions for agency is just one way to realize agency?

MULTIPLE REALIZATIONS OF AGENCY

→

EXTEND THE CONCEPTION OF AGENCY IN VARIOUS INTERESTING WAYS

DENKWERKSTATT BERLIN

II. Agency of children & non-human animals

SOPHISTICATED TERMINOLOGY OF PHILOSOPHY ALREADY REACHES ITS LIMITS WHEN IT COMES TO CHILDREN OR NON-HUMAN ANIMALS

❖ **developmental psychology & animal cognition demonstrate gradual appearances & multiple realizations of agency**

MINIMAL APPROACHES IN PHILOSOPHY EXTEND VARIOUS NOTIONS

(1) assuming multiple realizations & questioning demanding conditions of standard notions → not all conditions turn out to be necessary

(2) new set of minimal necessary conditions of socio-cognitive phenomena

minimal mindreading
(Butterfill & Apperly 2013)

shared intention light
(Pacherie 2013)

minimal action
(Strasser 2006)

minimal sense of commitment
(Michael et al. 2016)

(cf. Brownell 2011; Heyes 2014, 2015; Perler 2005; Premack/Woodruff 1978; Vesper et al. 2010; Warneken et al. 2006)

My strategy

BEING A PARTICIPANT IN A JOINT ACTION IS AN INDICATOR OF AGENCY

Minimal mindreading & young children

- capable of minimal mindreading
- sufficient to participate in joint actions

→ failing the (explicit) false-belief task and related tests in understanding others is not a reason to exclude them as participants in joint actions

Agency & artificial systems

- capable of being participants in asymmetric joint actions
- sufficient to ascribe minimal agency

→ lack of the biological make-up is not a reason to exclude them from agency

DENKWERKSTATT BERLIN

III. Joint actions first

INDIVIDUAL AGENCY IS NOT THE MOST PROMISING STARTING POINT

developmental psychology

- it seems that infants first interact with their caregivers before they begin to act individually

→ to become a full-fledged agent
– capable of individual action –
you must have had experiences with joint actions

- In the beginning, we need more-experienced agents who treat us as interaction partners. Step by step, we then grow into the role of being a participant in a joint action.

- *Some agents might only be able to act with more-experienced agents together.*

DENKWERKSTATT BERLIN

Joint actions everywhere

DENKWERKSTATT BERLIN

Bratman – joint actions

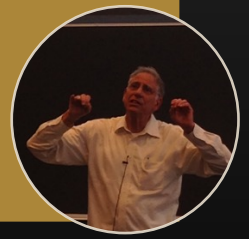
shared intentions & goals

specific belief state

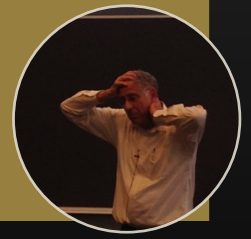
relation of interdependence & mutual responsiveness

common knowledge

mastery of mental concepts



sophisticated mentalization skills



(Bratman 2014)

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

ADULT & CHILD

ARTIFICIAL INTELLIGENCE

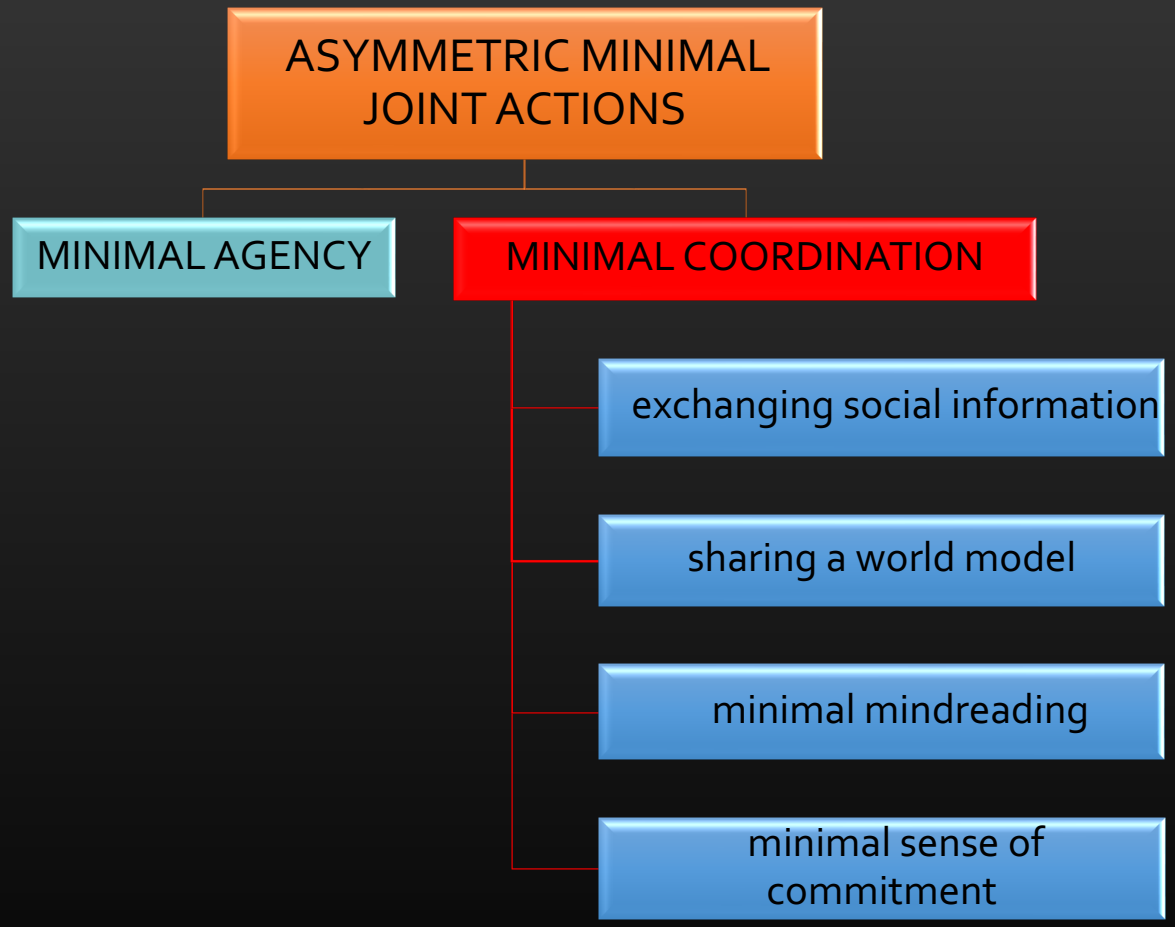
- joint action of human beings & artificial systems
- artificial systems =?= socially interacting entities

ROBOT & HUMAN

ARTIFICIAL AGENTS DO NOT HAVE TO FULFILL THE VERY SAME CONDITIONS AS HUMANS

DENKWERKSTATT BERLIN

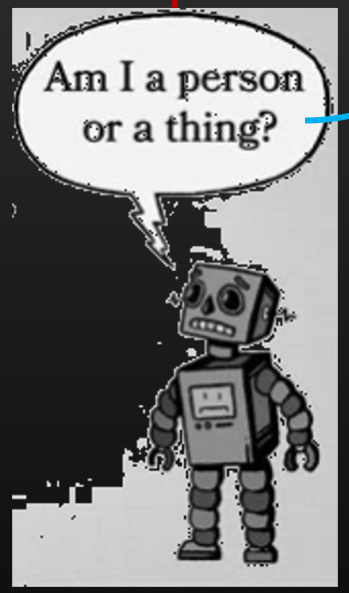
Set of minimal necessary conditions



IV. Agency of artificial systems

full-blown intentional systems → too inflationary

mere tools → too reductive



TOOL USE

INTERMEDIATE CATEGORY

SOCIAL INTERACTIONS

ASYMMETRIC JOINT ACTIONS

AI
- competent in means
- instrumental rationality

HUMAN
- competent in ends
- reflective rationality

The AI-Stance

DENKWERKSTATT BERLIN

Strasser, A. & Wilby, M. (2023). The AI-Stance: Crossing the Terra Incognita of Human-Machine Interactions?

Stance epistemology

The AI-Stance as a framework for understanding the intermediate category

inspired by Daniel Dennett's stance epistemology

- ***The physical stance*** is appropriate for explaining physical objects in general.
- ***The design stance*** is useful for explaining interactions with objects that fulfill a fixed purpose.
- ***The intentional stance*** is helpful in front of objects that operate according to intentional, belief/desire explanations.

The design stance towards computers can allow one to "predict its behavior with great accuracy and reliability."

The intentional stance can be appropriate to computers because this stance equips us with successful anticipations of their behavior.

- reducing all interactions with artificial systems to mere tool-use

certain artificial systems act according to intentional and rational patterns

- not appropriate for explaining & anticipating the contributions artificial agents can make in an asymmetric joint action

- ***The AI-Stance*** to explain the contribution of artificial systems in asymmetric joint actions.

DENKWERKSTATT BERLIN

The AI-Stance
capturing the more asymmetric conception of joint actions

➢ **MEANS-ENDS REASONING**

SYMMETRIC JOINT ACTION
- both have a goal in mind / various means by which a goal can be realized

➢ THEY SHARE THE SAME TOKEN END BUT MIGHT CONTRIBUTE DIFFERENT MEANS TOWARDS THAT END

expertise in ends & means tends to be matched

ASYMMETRIC JOINT ACTIONS

Expert in means

Expert in ends

almost total mismatch is possible

- artificial systems might be highly expert in means
- with respect to ends, they are rather silent
- the ends are programmed in or stipulated by the human interactant

MEANS: HIGHLY COMPETENT
artificial systems might be highly expert in means

ENDS:
NO COMPETENCE
artificial systems are silent

DENKWERKSTATT BERLIN

Instrumental & reflective rationality

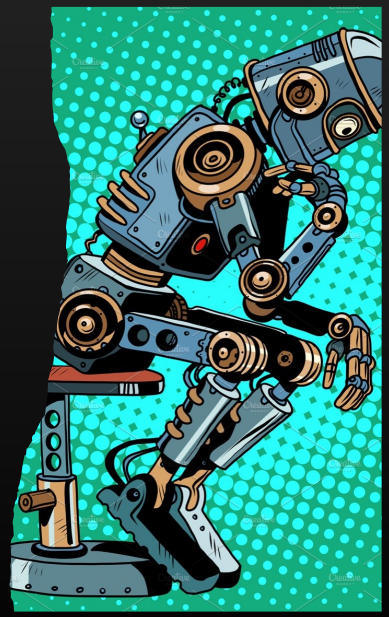
instrumental rationality
adopting suitable means to ends that are already fixed
EXPERT IN MEANS

reflective rationality
choosing, evaluating and reconsidering ends themselves
EXPERT IN ENDS

not yet available to artificial systems
- because it would require AGI

Reflective rationality barely figures within the goals of 'current AI research.'
(Stuart Russell 2020)

THE END IS NOT CHOSEN BY THE MACHINE



DENKWERKSTATT BERLIN

Asymmetric joint actions with machines

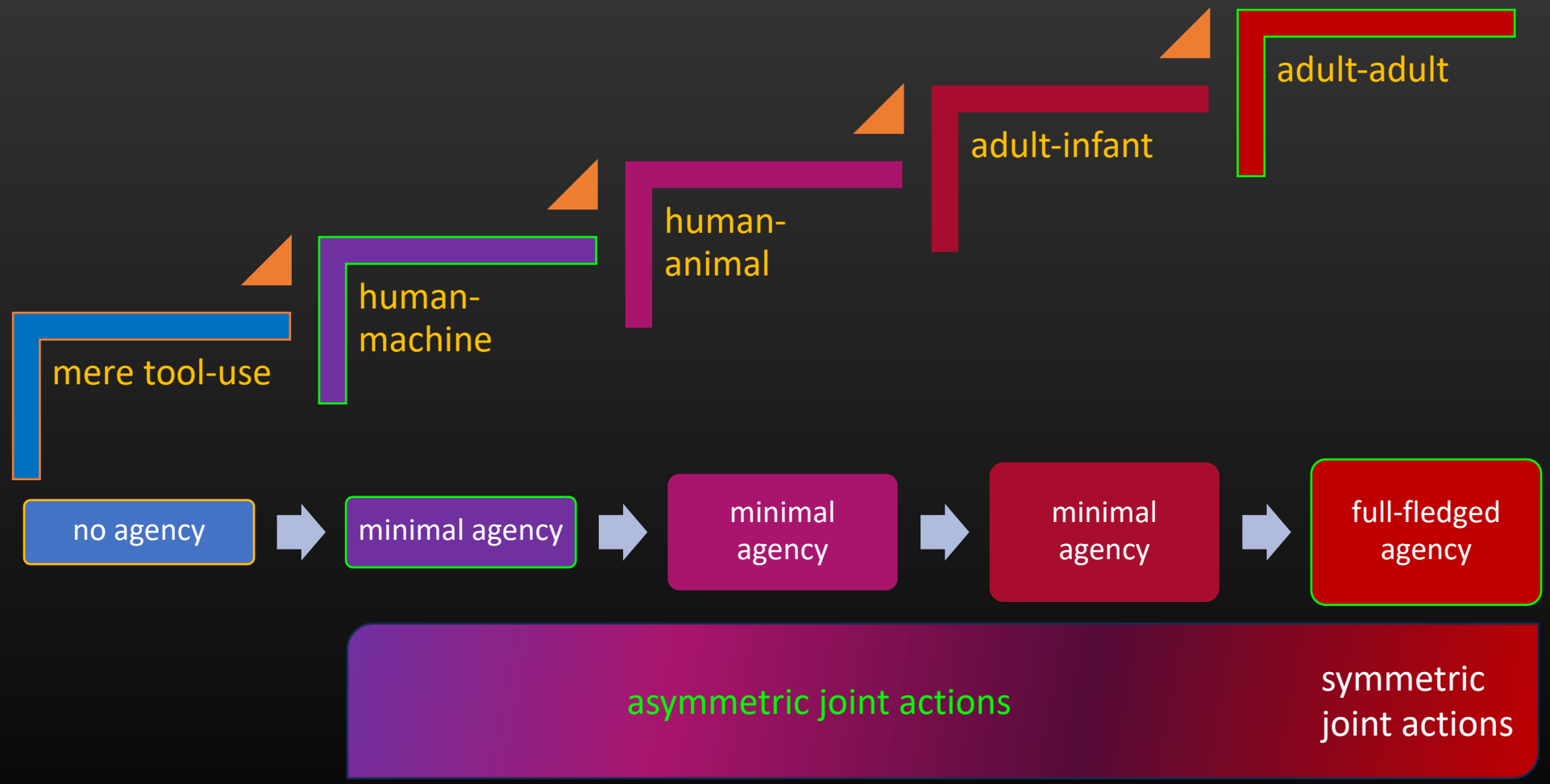
IF ONE IS INTERACTING WITH A MACHINE, THEN, ONE IS INTERACTING WITH SOMETHING THAT IS POTENTIALLY CAPABLE OF **HIGHLY SOPHISTICATED INSTRUMENTAL RATIONALITY** BUT IS NOT AN EXPERT OF **REFLECTIVE RATIONALITY.**

To take the AI-stance is to be prepared to treat the AI-system as

- **FULLY RATIONAL IN TERMS OF INSTRUMENTAL RATIONALITY**
- **ARATIONAL IN TERMS OF REFLECTIVE RATIONALITY**

DENKWERKSTATT BERLIN

Conclusion


mere tool-use
human-machine
human-animal
adult-infant
adult-adult
no agency
minimal agency
minimal agency
minimal agency
full-fledged agency
asymmetric joint actions
symmetric joint actions

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

Josh Rust

Mike Wilby

Eric Schwitzgebel

Thank You !

DENKWERKSTATT BERLIN

References

Butterfill S, Apperly I. (2013). How to construct a minimal theory of mind. *Mind and Language,* 28 (5), 606-637.

Davidson D. (1963). Actions, reasons and causes. *Journal of Philosophy*, 60(23): 685-99.

(1971). Agency. In: Binkley R, Bronaugh R, Marras, A., (eds.) *Agent, action and reason.* University of Toronto Press, 3-37.

(1980). *Essays on actions and events.* Oxford University Press.

(1984). *Inquiries into truth and interpretation.* Oxford University Press.

(1982). Rational animals. *Dialectica*, 36: 317-28.

(2001). *Subjective, intersubjective, objective.* Oxford University Press.

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

Heyes C. (2015). Animal mindreading: what's the problem? *Psychonomic Bulletin & Review;* 22 (2), 313-327.

(2014). False belief in infancy: a fresh look. *Developmental Science,* 17(5), 647-659.

Hurley, S. (2003). Animal Action in the Space of Reasons. *Mind & Language*, 18: 231-257.

Michael J, Sebanz N, Knoblich G. (2016). The Sense of Commitment: A Minimal Approach. *Frontiers in Psychology*, 6, 1968.

Pacherie E. (2013). Intentional joint agency: Shared intention lite. *Synthese,* 190 (10), 1817-1839.

Perler D, Wild M. (2005). *Der Geist der Tiere – Philosophische Texte zu einer aktuellen Diskussion.* Frankfurt: Suhrkamp.

Perner J. (1991). *Understanding the representational mind.* Cambridge, MA: MIT Press.

References

Premack D, Woodruff G. (1978). Does the chimpanzee have a theory of mind? *Behavioral Brain Sciences,* 1, 515-526.

Russell S. (2020). Artificial intelligence: A binary approach. *In: Liao SM (ed.), Ethics of artificial intelligence.* Oxford University Press, 327.

Searle J. (1984). *Minds, Brains and Science: The 1984 Reith lectures.* London: British Broadcasting Corporation.

Sterelny K. (2014). *The evolved apprentice: How evolution made humans unique.* Cambridge, MA; MIT Press.

Steward, H. (2009). Animal Agency. *Inquiry*, 52:3, 217-231.

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

(2015). Can artificial systems be part of a collective action? In: Misselhorn C (ed.), *Collective Agency and Cooperation in Natural and Artificial Systems. Explanation, Implementation and Simulation.* Berlin: Springer; Philosophical Studies Series, 122, 205-218. doi: 10.1007/978-3-319-15515-9_11.

(2020). From tools to social agents. *Rivista Italiana di Filosofia del Linguaggio;* 14 (2), 76-87, doi: 10.4396/AISB201907.

(2022). From tool use to social interactions. In: Janina Loh and Wulf Loh (eds.), *Social Robotics and the Good Life. The Normative Side of Forming Emotional Bonds With Robots.* transcript publishing.

Strasser, A. & Wilby, M. (2023). The AI-Stance: Crossing the Terra Incognita of Human-Machine Interactions? In R. Hakli, P. Mäkelä, J. Seibt (eds.), *Social Robots in Social Institutions. Proceedings of Robophilosophy 2022.* Series Frontiers of AI and Its Applications, vol. 366. IOS Press, Amsterdam.

Tomasello M. (2008). *Origins of human communication.* MIT Press.

Vesper C, Butterfill S, Knoblich G, Sebanz, N. (2010). A Minimal Architecture for Joint Action. *Neural Networks*; 23, 998-1003.

Warneken F, Chen F, Tomasello M (2006). Cooperative activities in young children and chimpanzees. *Child Dev.*; 77, 640-663.

Wellman H, Cross D, Watson J. (2001). Meta-Analysis of Theory of Mind Development: The Truth About False-Belief. *Child Dev.;* 72 (3): 655-84.