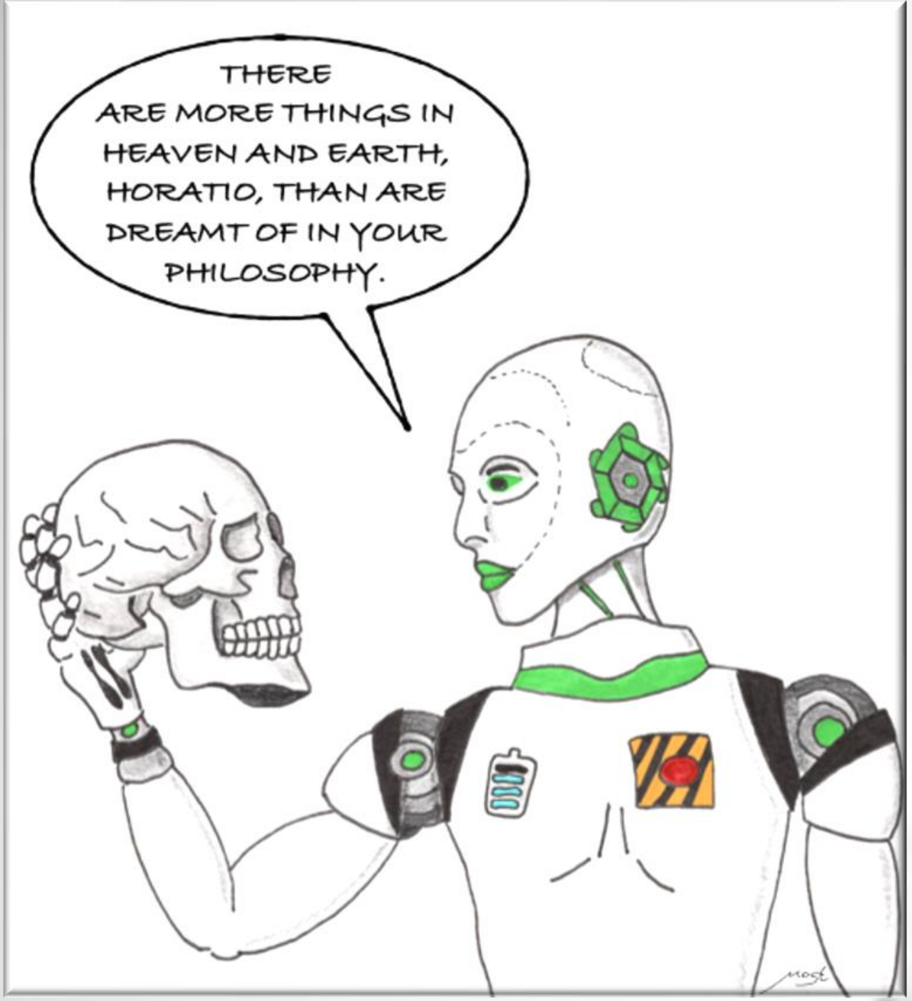

THERE ARE MORE THINGS IN HEAVEN AND EARTH, HORATIO, THAN ARE DREAMT OF IN YOUR PHILOSOPHY.

Artist: Moritz Strasser

INBETWEENISM

NEITHER/NOR –

TOWARDS INBETWEENISM

(Chapter 3)

ANNA STRASSER (DenkWerkstatt Berlin, Germany)

Things don’t dichotomize

PHILOSOPHY POSES TOO DEMANDING CONDITIONS

Artist: Lorin Strasser

- ➢ 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
- ➢ explore how one could expand or adopt the sophisticated terminology of philosophy to capture phenomena one finds in developmental psychology, animal cognition, and AI

➢ GRADUAL APPROACHES & MINIMAL NOTIONS

Philosophy & Cognitive Science

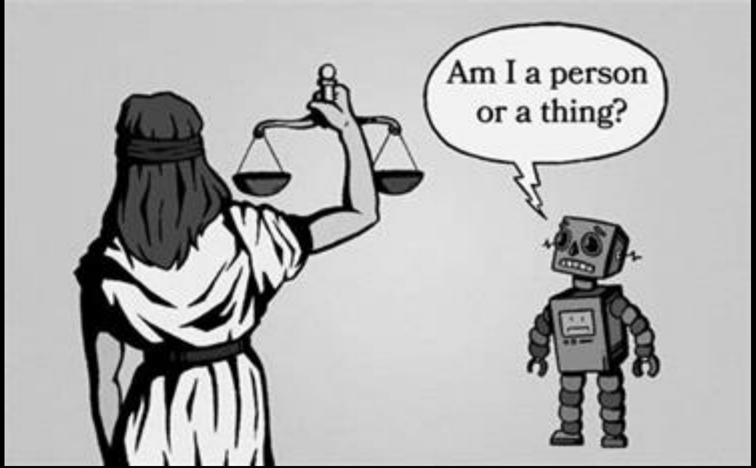
Anna Strasser

Kognition künstlicher Systeme

ontos verlag

(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

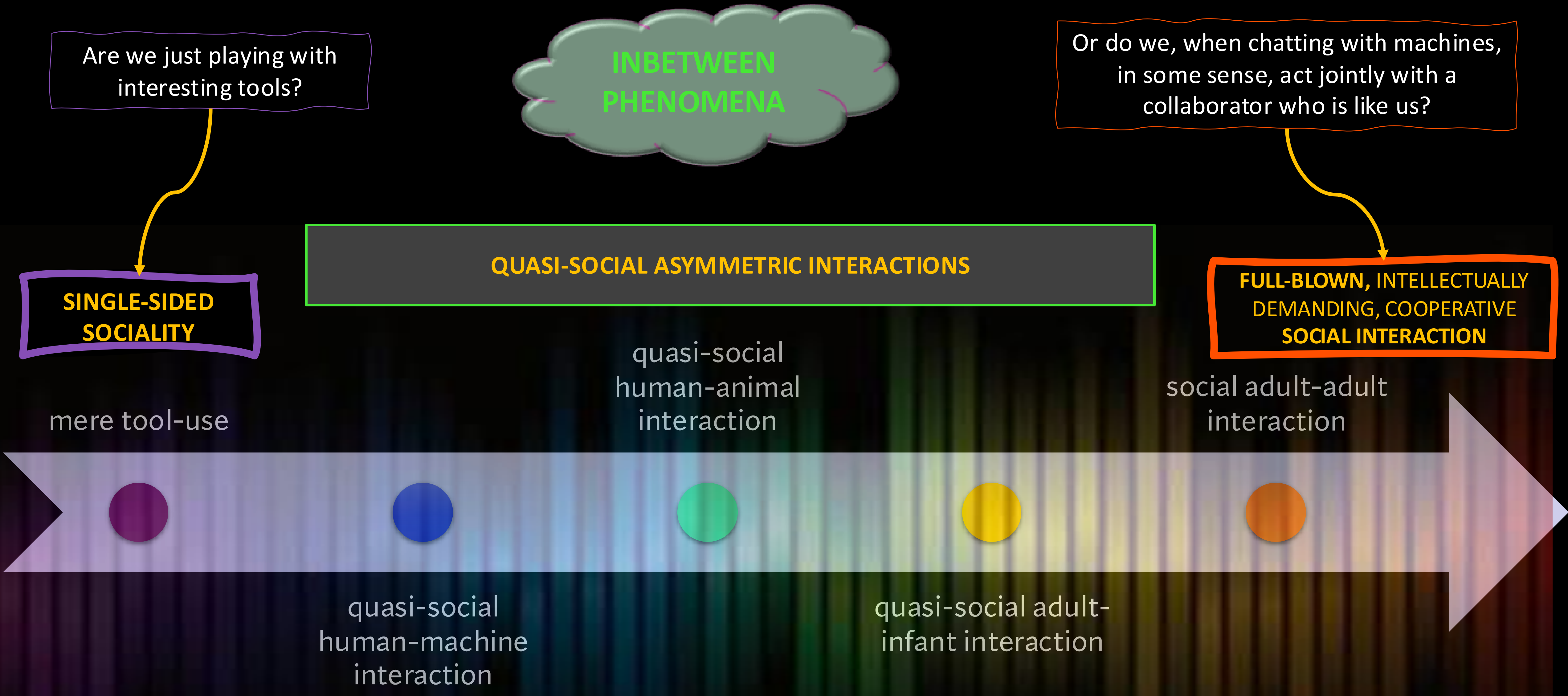
WE CANNOT REDUCE ALL OF OUR INTERACTIONS WITH LLMS TO MERE TOOL USE

“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.”

→ RETHINK OUR CONCEPTUAL FRAMEWORK

which so clearly distinguishes *between tools as inanimate things* and *humans as social, rational, and moral interaction partners*

A multidimensional spectrum of social interactions



Are we just playing with interesting tools?

INBETWEEN PHENOMENA

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

SINGLE-SIDED SOCIALITY

QUASI-SOCIAL ASYMMETRIC INTERACTIONS

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

mere tool-use

quasi-social human-machine interaction

quasi-social human-animal interaction

quasi-social adult-infant interaction

social adult-adult interaction

What happened so far?

Chapter 1

- analysis of the relevant concepts (agency, moral agency, moral patiency)
  → restrictive use of these concepts assumes that only living beings can qualify

- several motivations to question the dichotomy between animate & inanimate (respectively, mere tool use & full-fledged social interactions)

INSISTING ON THIS DICHOTOMY, ONE CAN ONLY TAKE ONE OF TWO EXTREME POSITIONS:

- *Hard-core instrumentalist:*
  excluding the possibility that any artificial system could have a social status in an HMI

- *In-expectation of AGI view:*
  whole demanding package of conditions that we require from humans in terms of agency, moral agency and moral patiency can in principle also be fulfilled by sophisticated machines
  → artificial life

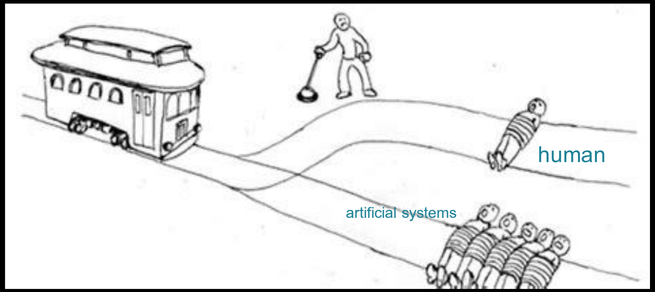
What happened so far?

Chapter 2: both options are not very attractive when it comes to ethical questions

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

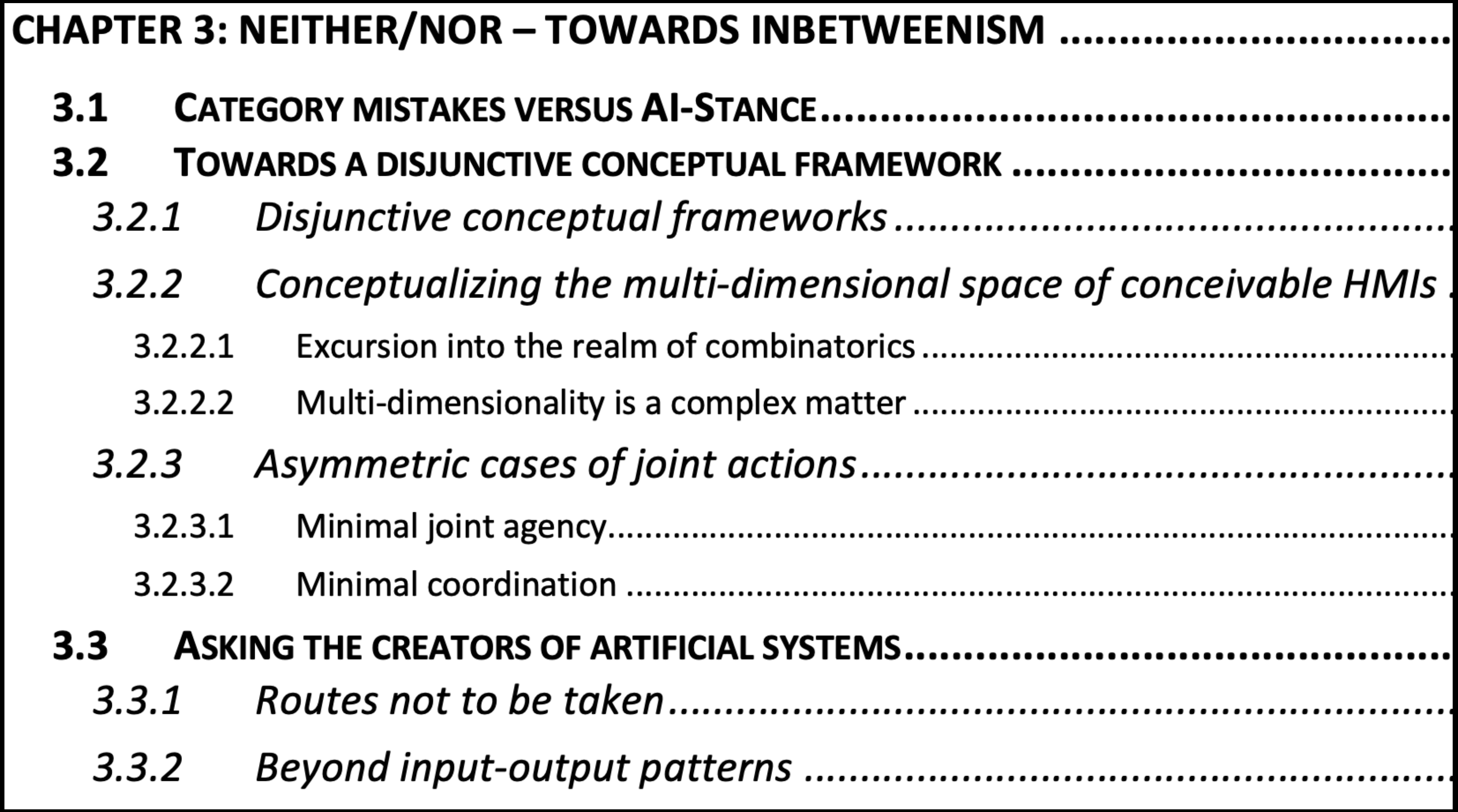
- no straight-forward reasons to allow our interactions with artificial systems to be guided by moral or social norms



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

Chapter 3 NEITHER/NOR – TOWARDS INBETWEENISM



3.1. Category mistakes versus AI-Stance

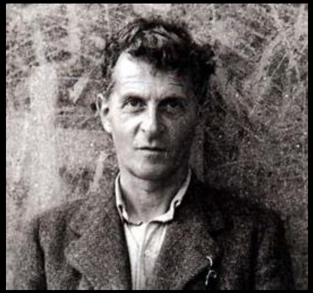
THINGS DON'T DICHOTOMIZE

skip

- partial anthropomorphizing that amounts to ascribing some of the socio-cognitive abilities that normally are only ascribed to humans is not necessarily misleading
  - according to the AI-Stance that I developed together with Michael Wilby, it is sufficient to ascribe instrumental rationality to the artificial interaction partners

3.2. Towards a disjunctive conceptual framework

HOW TO CHARACTERIZE THE MANY DIFFERENT INSTANCES IN A MULTI-DIMENSIONAL SPECTRUM OF SOCIAL INTERACTIONS



Wittgenstein, Ludwig. 2009. *Philosophical investigations*.

INSTANCES STAND IN A RELATION OF FAMILY RESEMBLANCE

ALLOWING MULTIPLE REALIZATION

- advocate for a disjunctive conceptual framework that 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

ACKNOWLEDGING A GRADUAL APPROACH TOWARDS ABILITIES THAT ARE REQUIRED

- expand the field of application of various notions describing required abilities
- follow the strategy of minimal approaches (question the necessity of some conditions that come with the standard notions from philosophy and allow for a less strong manifestation of required abilities)

minimal approaches

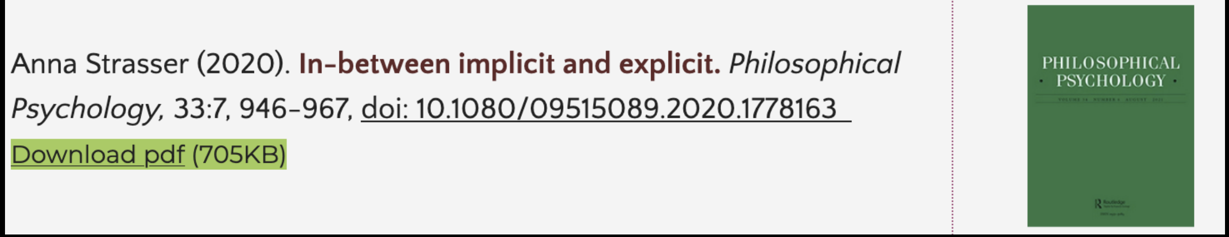
Stephen Butterfill & Ian Apperly (2013): minimal mindreading | John Michael et al. (2016): minimal sense of Commitment | Elisabeth Pacherie (2013): shared intention lite | Anna Strasser (2006): minimal action

3.2.1. Disjunctive conceptual frameworks

INSIGHTS FROM OF ANOTHER DICHOTOMY, NAMELY THE ONE BETWEEN IMPLICIT AND EXPLICIT COGNITIVE PROCESSES

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



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

Download pdf (705KB)

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.

	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

3.2.2. 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

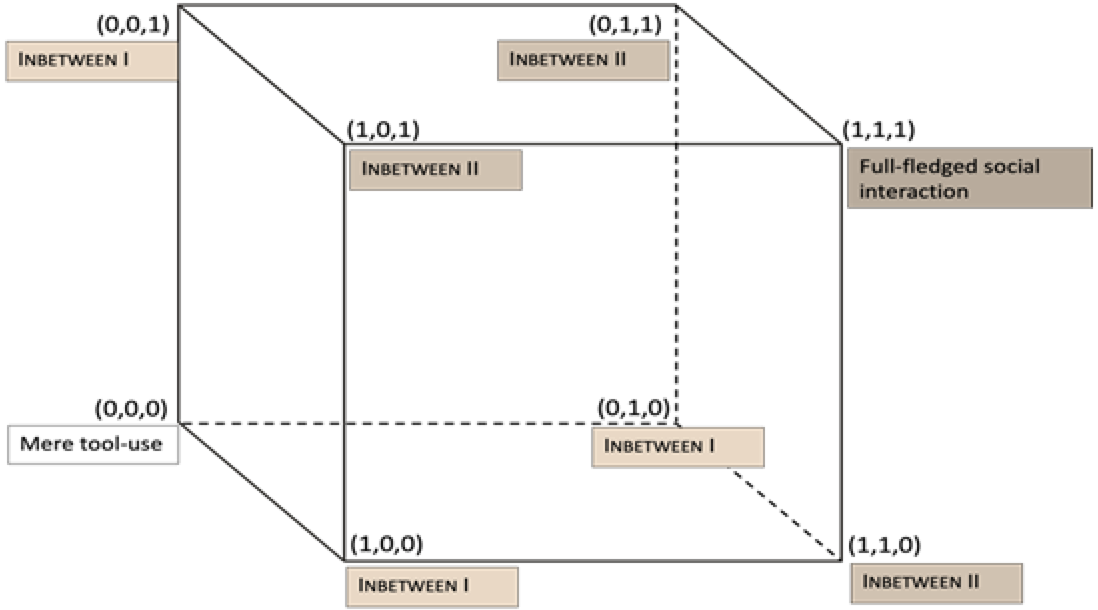
➢ be prepared for the possibility that there will be no clear-cut criteria to establish a sharp border

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.

3.2.2.1. Excursion into the realm of combinatorics

When asking how to order all conceivable instances in a multi-dimensional spectrum, we will see that this question cannot always be answered

skip



3.2.2.2. 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

- THEN 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, of course, not emerge in an instant

- *be that developmentally in humans,*
- *phylogenetically in animal evolution, or*
- *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.

3.2.3. 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

- 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
LLM & HUMAN

DISTINCT TYPES OF ASYMMETRIC JOINT ACTIONS ARE CONCEIVABLE

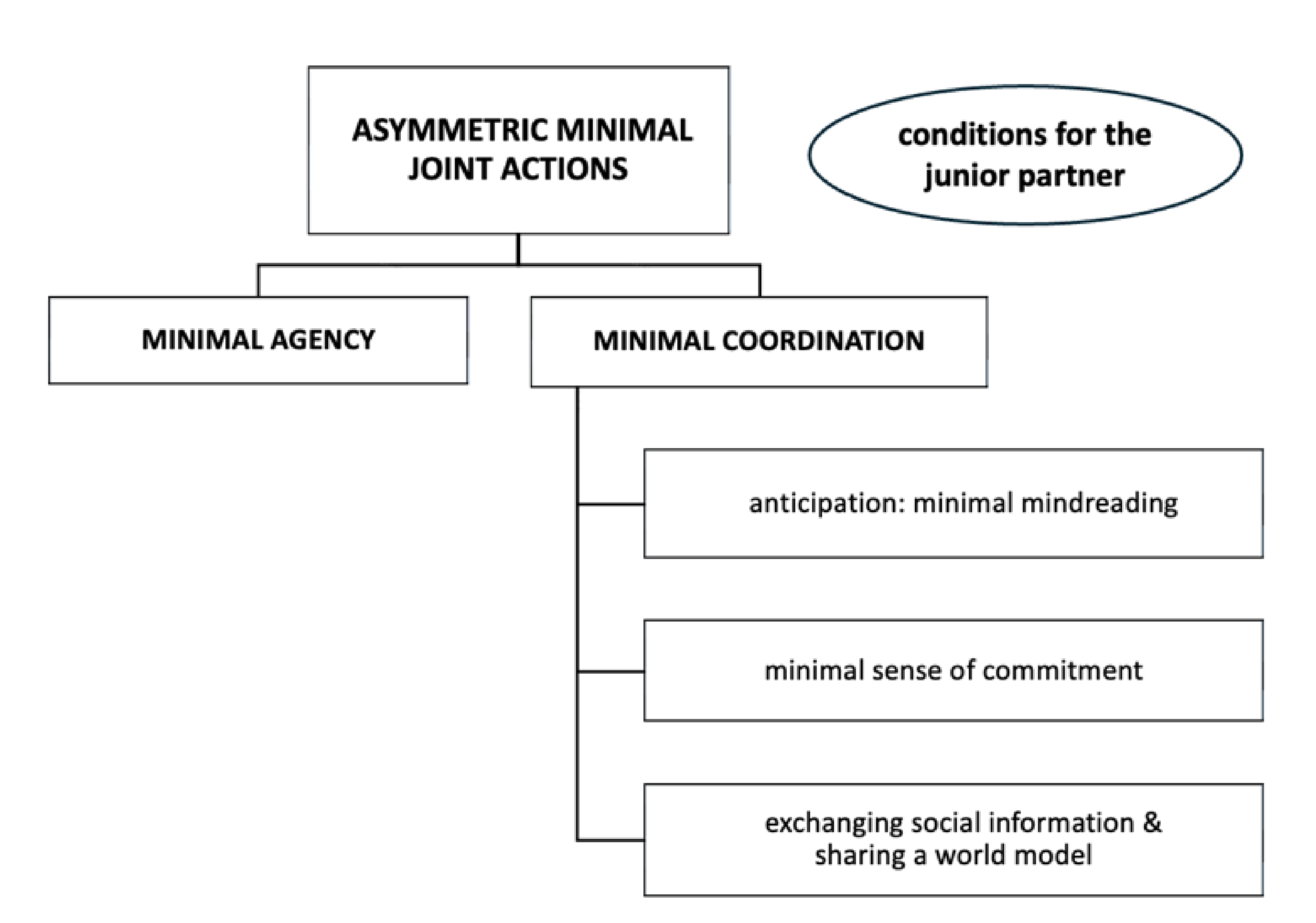
whereby each type differs with respect to the proposed set of conditions

To avoid any misunderstandings, I want to emphasize that I do not equate interactions with children with interactions with artificial systems – they only share the characteristic of both being asymmetric.

3.2.3. 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?



ASYMMETRIC MINIMAL JOINT ACTIONS

conditions for the junior partner

MINIMAL AGENCY

MINIMAL COORDINATION

anticipation: minimal mindreading

minimal sense of commitment

exchanging social information & sharing a world model

REQUIREMENTS FOR AGENCY & OTHER SOCIO-COGNITIVE ABILITIES

THAT CAN ENSURE THAT ARTIFICIAL AGENTS HAVE

SUFFICIENT ABILITIES TO QUALIFY AS QUASI-SOCIAL INTERACTION PARTNERS

3.2.3.1. Minimal joint agency

With reference to my dissertation *Kognition künstlicher Systeme* I pose several conditions:

Artificial systems in question have to

(1) be cognitive systems with a flexible coupling between input & output which implies a learning ability and a degree of autonomy by which they exhibit can goal-oriented behavior

(2) be capable of action in our world
→ they need the ability to take in relevant information and represent it in a world model
→ flexibility in the information processing procedures should enable them to adapt to environmental change and acquire knowledge in relation to an action goal

(3) have effectors that can trigger changes in the environment

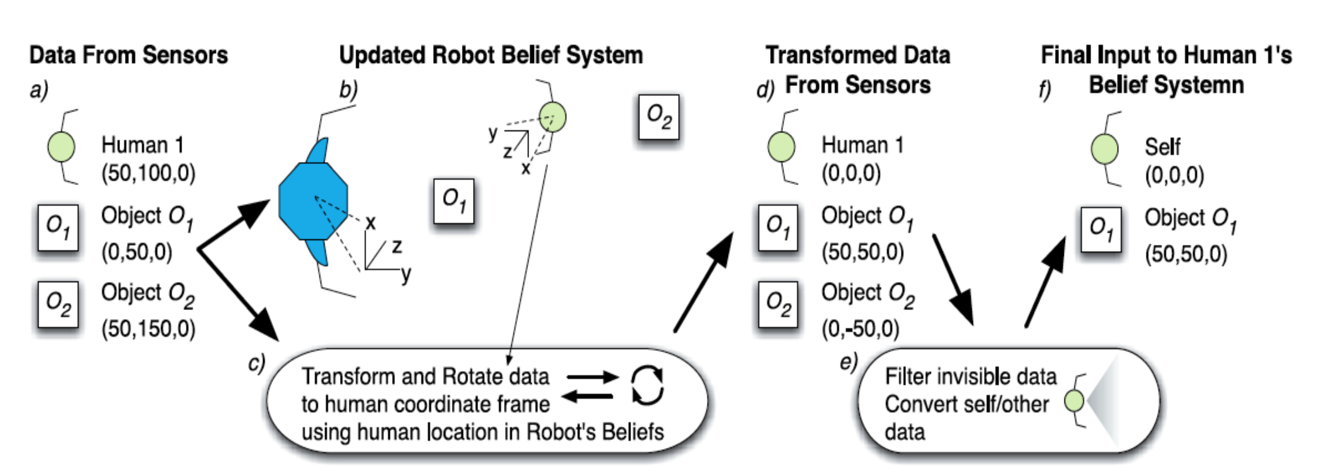
(4) demonstrate their ability to act through adapting to a dynamic environment

Framing the slogan 'joint action first,' in the first chapter of this book, I argued in addition for the claim that if we are asking for agentive properties in HMIs, we do not necessarily have to assume individual agency from each potential interaction partners – joint agency is sufficient.

3.2.3.2. Minimal coordination

skip

MODELLING MENTAL STATES WITH RESPECT TO THE PERSPECTIVE OF THE HUMAN COUNTERPART



1. infer from their perception of the physical world to what a human counterpart can see or cannot see
2. infer that perspective of the human will guide future actions of the human

→ some cases of minimal mindreading can be achieved by artificial agents

LLMs live NOT

in our social, physical world

are not embodied

But they may play a role in our world of language games.

(Gray & Breazal, 2014)

3.3. Asking the creators of artificial systems

3.3.1. Routes not to be taken

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

- 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

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

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

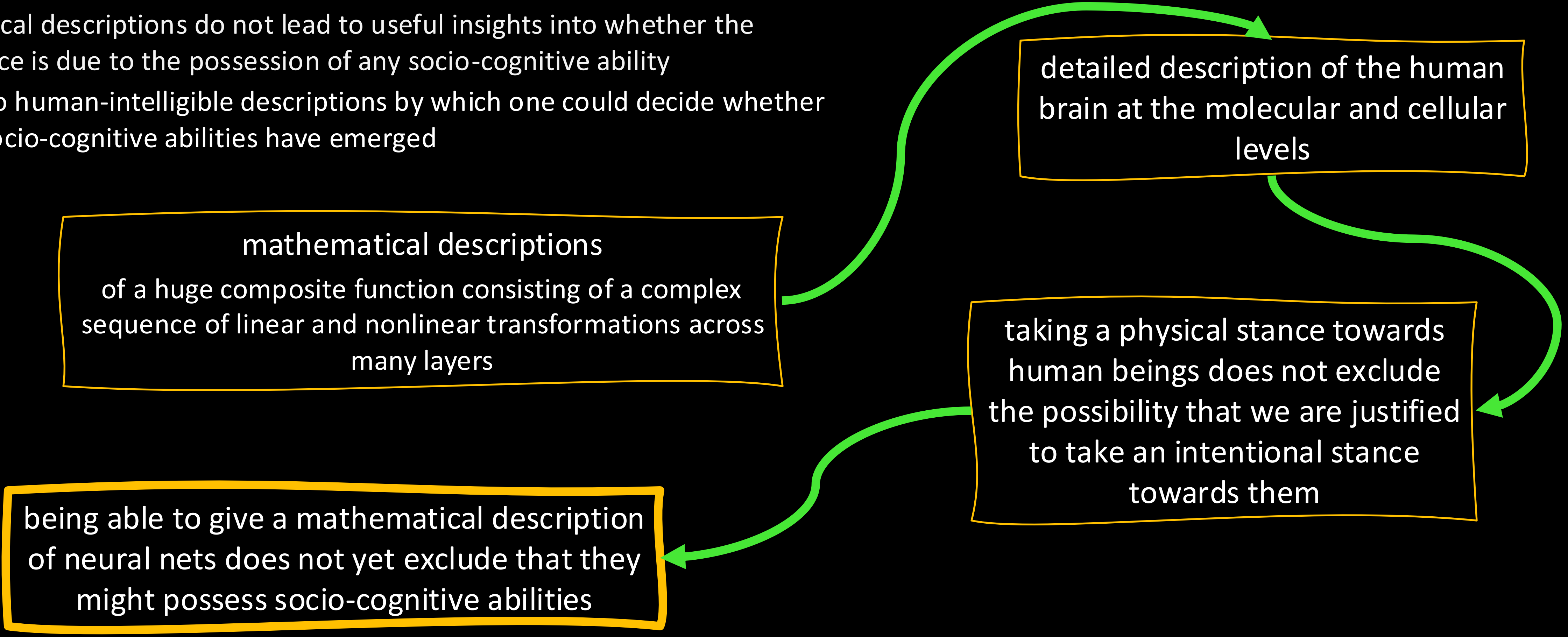
3.3. Asking the creators of artificial systems

3.3.2. BEYOND INPUT-OUTPUT PATTERNS

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

mathematical descriptions do not lead to 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



mathematical descriptions

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

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

taking a physical stance towards human beings does not exclude the possibility that we are justified to take 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 arguments 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

3.3. Asking the creators of artificial systems

3.3.2. Beyond input-output patterns

skip

PROBING, ATTRIBUTION, CAUSAL INTERVENTION

INTERPRETABILITY TECHNIQUES

investigating the inner structure of neural networks

- aim to uncover the causal mechanisms underlying LLMs' performance at a higher level
- asking whether LLMs
  - represent information
  - operate on representations
  - have activation patterns that realize socio-cognitive abilities follows

probing

- exploring what is encoded in a neural network. From this, one can then make statements that certain information is likely to be represented in their activation pattern, however, this does not yet provide any information as to whether these representations are used when the model solves a task.

attribution methods

- explore which parts of the input data (the prompts provided by the human interaction partner) a model relies on most for their outputs

causal intervention methods

- determine the causal role played by a representation in the processing of a model
  - models are changed in various ways, and it is examined whether the intervention changes the predictions (the outputs) of the model in a systematic way
    → hypotheses regarding the processing are tested, e.g., whether a model performs a systematic calculation to solve the task or whether a system has something like a world mode

Conclusion

- To conclude, one can say that the ascription of properties and socio-cognitive abilities to artificial systems cannot be clarified by computer science alone.

- However, 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.

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 !