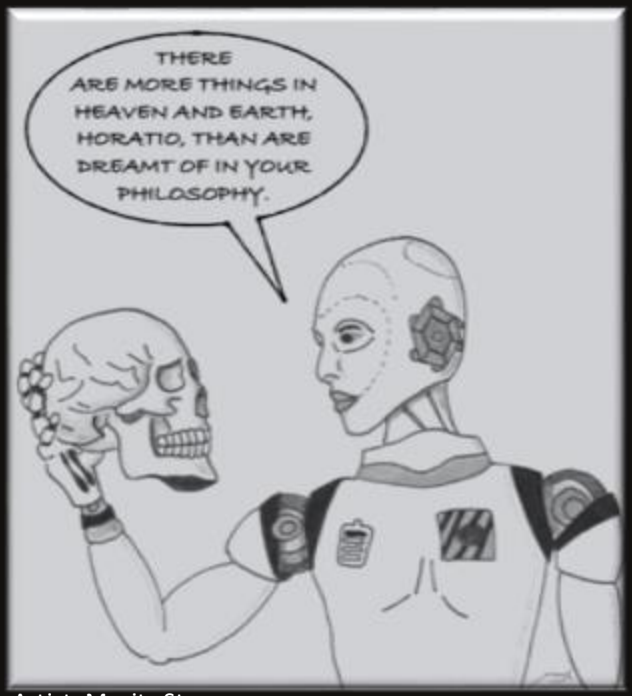
ANNA STRASSER

(DenkWerkstatt Berlin, Germany)

NEITHER/NOR – TOWARDS INBETWEENISM

WHAT TO DO IF THINGS DON'T DICHOTOMIZE



Artist: Moritz Strasser

Overview

CONCEPTUAL PROBLEM

Things don't dichotomize

- A multidimensional spectrum of social interactions
- Before developing a disjunctive conceptual framework
- Why current ethical theories are outdated

DISJUNCTIVE CONCEPTUAL FRAMEWORK

Towards a disjunctive conceptual framework

- Other disjunctive conceptual frameworks
- Conceptualizing the multi-dimensional space of conceivable HMIs
- Excursion into the realm of combinatorics
- Multi-dimensionality is a complex matter

Asymmetric distribution of abilities

- Asymmetric cases of joint actions
- Minimal joint agency | Minimal coordination

ASCRIPTION PROBLEM

Another severe obstacle

- Asking the creators of artificial systems
- Beyond input-output patterns
- Interpretability techniques

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

Things don't dichotomize

CONCEPTUAL FRAMEWORKS SHOULD NOT FORCE US TO DICHOTOMIZE

INTELLECTUALIST APPROACHES IN PHILOSOPHY POSE TOO DEMANDING CONDITIONS

- 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

(Strasser, 2006)

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

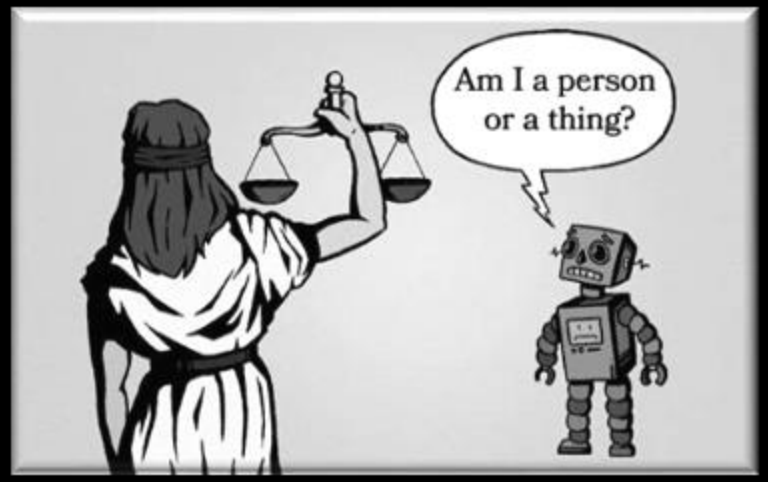
Artist: Lorin Strasser

GRADUAL APPROACHES

- disjunctive conceptual framework enriched with minimal notions that can capture all kinds of inbetween phenomena

A conceptual problem

WE DO NOT HAVE THE RIGHT KIND OF NOTIONS TO DESCRIBE CERTAIN PHENOMENA ADEQUATELY



- 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 our interactions with LLMs are properly asocial (just tool-use or self-talk) nor quite right to say that our interactions with LLMs are properly social. Neither standard philosophical theorizing nor dichotomous ordinary concepts enable us to think well about these in-between phenomena."

Strasser & Schwitzgebel 2024, 197

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

A multidimensional spectrum of social interactions

CAN BE CONCEPTUALIZED WITH THE HELP OF A DISJUNCTIVE CONCEPTUAL FRAMEWORK

Are we just playing with interesting tools?

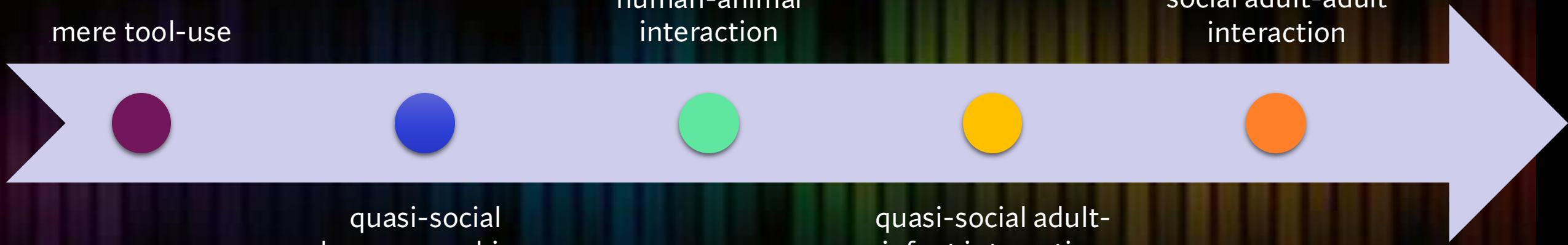
INBETWEEN PHENOMENA

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

QUASI-SOCIAL ASYMMETRIC INTERACTIONS

SINGLE-SIDED SOCIALITY

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



Before developing a disjunctive conceptual framework

CHAPTER 1: 1. ETHICS AS A GUIDE FOR MORAL AGENTS

- analysis of relevant concepts (agency, moral agency, moral patiency) to describe the role of artificial systems in HMIs
  → restrictive use of these concepts assumes that only living beings can qualify

Why we should 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

Why current ethical theories are outdated

Chapter 2: Challenges posed by new AI technology

BOTH OPTIONS ARE NOT VERY ATTRACTIVE WHEN IT COMES TO ETHICAL QUESTIONS

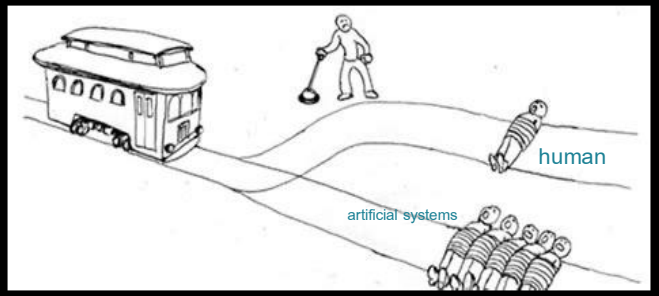
Hard-core instrumentalists

- 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

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.*

Acknowledging a gradual approach towards required abilities

- expand the range 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

Instances stand in a relation of family resemblance

- allow multiple realization

Advocate for 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

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

Other 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)

EITHER/OR DISTINCTION BETWEEN EXPLICIT & IMPLICIT PROCESSES

- ➢ 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

1. manifestations of conditions in various strengths
2. less demanding conditions can prove sufficient
3. 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

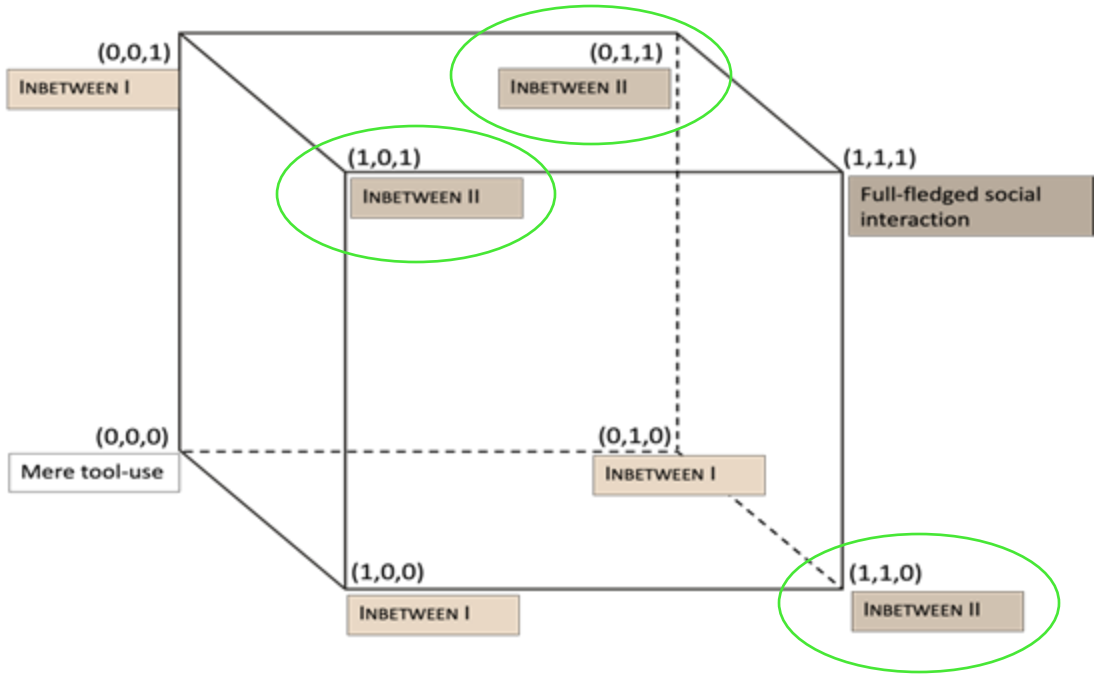
- no concrete hypothesis which of the many conceivable combinations of socio-cognitive abilities finally turn out to be sufficient
- advocating a gradual approach, the question of resemblance is a matter of degree

➢ 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.

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



Who is more social?

It is unclear which of the three combinations of fulfilled conditions proves to be more social.

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.

- QUASI-SOCIALITY EXISTS ON A COMPLEX SPECTRUM

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 & children
- children = socially interacting beings

ARTIFICIAL INTELLIGENCE

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

ADULT & CHILD

ROBOT & HUMAN
LLM & HUMAN

DISTINCT TYPES OF ASYMMETRIC JOINT ACTIONS ARE CONCEIVABLE

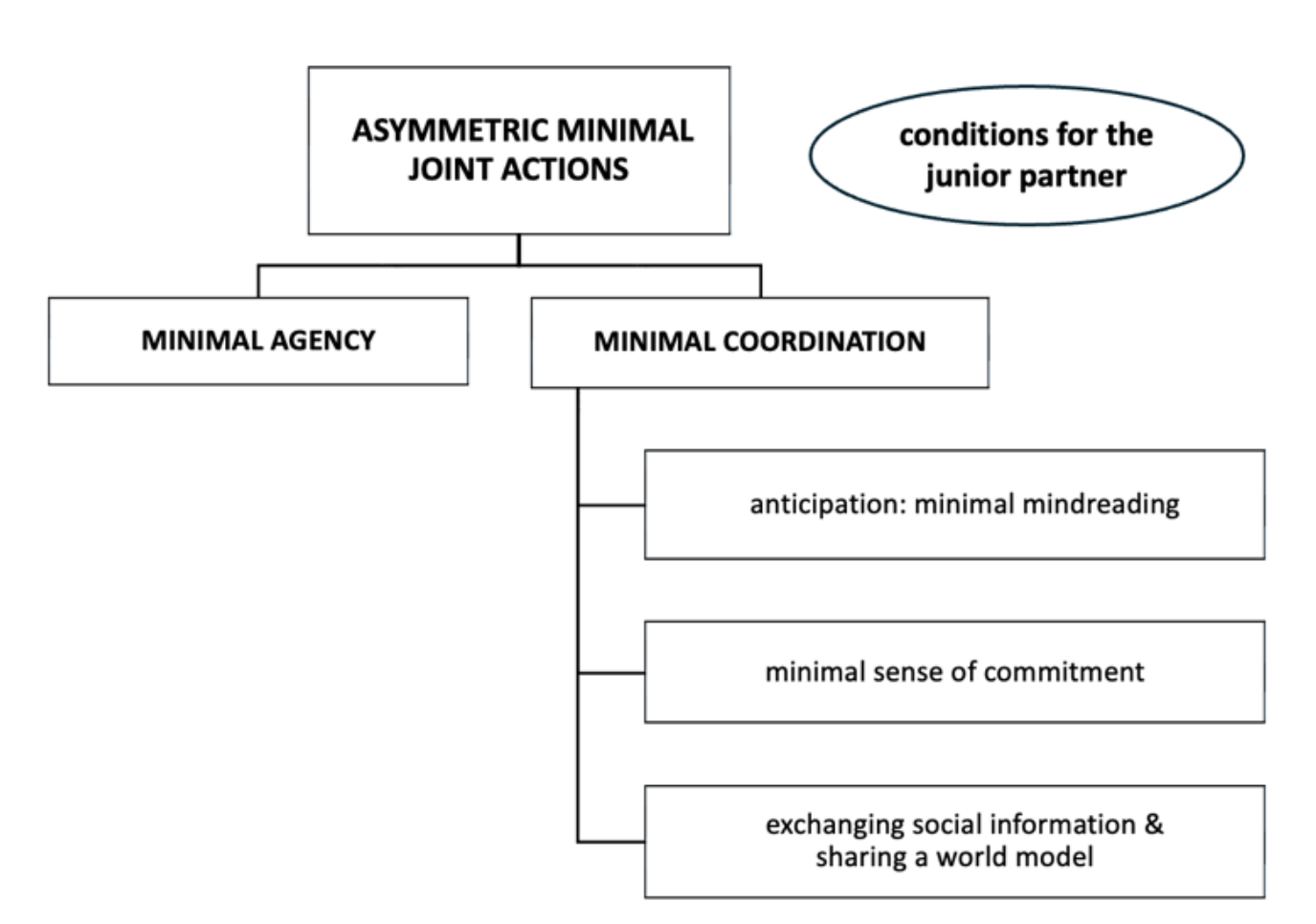
whereby each type differs with respect to the proposed set of conditions

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.

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

Minimal joint agency

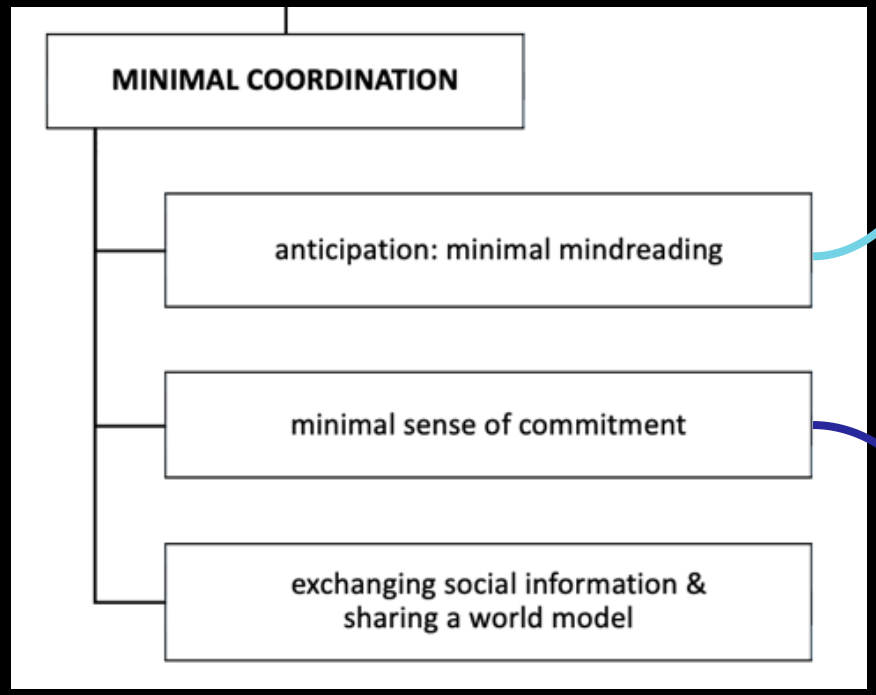
With reference to my dissertation *Kognition künstlicher Systeme,* I pose several conditions:

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 can exhibit 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 by 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.

Minimal coordination



anticipate what the other agent will do next

form expectations and motivations with respect to your counterpart

utilize the notion of minimal mindreading (Butterfill & Apperley, 2013)

- notion is a suitable starting point
  - as they claim that underlying processing are implicit, nonverbal, automatic, and based on unconscious reasoning


utilize the notion of a minimal sense of commitment (Michael et al., 2016)

- components (expectation or motivation) of a standard commitment can be disassociated
- single occurrence of just one component can be treated as a sufficient condition



Another severe obstacle

If you look at the debate about justified attributions of properties and abilities to artificial systems, irreconcilable positions clash.

As far as I can tell, there is no good chance that the various parties will agree in the foreseeable future on what properties and abilities these new types of smart machines ultimately have.

Asking the creators of artificial systems

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
(Wittgenstein / Kripke)

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

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

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

mathematical descriptions

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

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

Interpretability techniques

AIM TO UNCOVER THE CAUSAL MECHANISMS UNDERLYING LLMs' PERFORMANCE AT A HIGHER LEVEL

investigating the inner structure of neural networks by asking whether LLMs

- represent information
- operate on representations
- have activation patterns that realize socio-cognitive abilities

probing

- exploring what is encoded in a neural network.

➢ statements that certain information is likely to be represented in their activation pattern

BUT does not yet provide 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

THOSE TECHNIQUES PRESUPPOSE OPERATIONALIZABLE THEORIES

BUT this is a problem because as we do not yet have mainstream theories with respect to all socio-cognitive abilities.

AND we will have to wait until those techniques can also be applied to large language models as up to now they are practiced with toy models.

Conclusion

CONCEPTUAL PROBLEM

Certain HMIs are INBETWEEN phenomena as they are neither mere tool use not full-fledged social interactions.
To describe them adequately we need a new conceptual framework that does not force us to dichotomize.
Otherwise, we could only choose between hard-core instrumentalism or the In expectation of AGI view.

DISJUNCTIVE CONCEPTUAL FRAMEWORK

Advocating a gradual approach, I suggest that a disjunctive framework can capture a multi-dimensional spectrum of quasi-social interactions that includes asymmetric interactions in which the required conditions of involved participant can vary. All instances stand in a relation of family resemblance.

ASCRIPTION PROBLEM

It is a controversial debate of how one can argue for justified ascriptions of conditions that are required by the suggested framework.
I demonstrated 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.

Conclusion

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 !