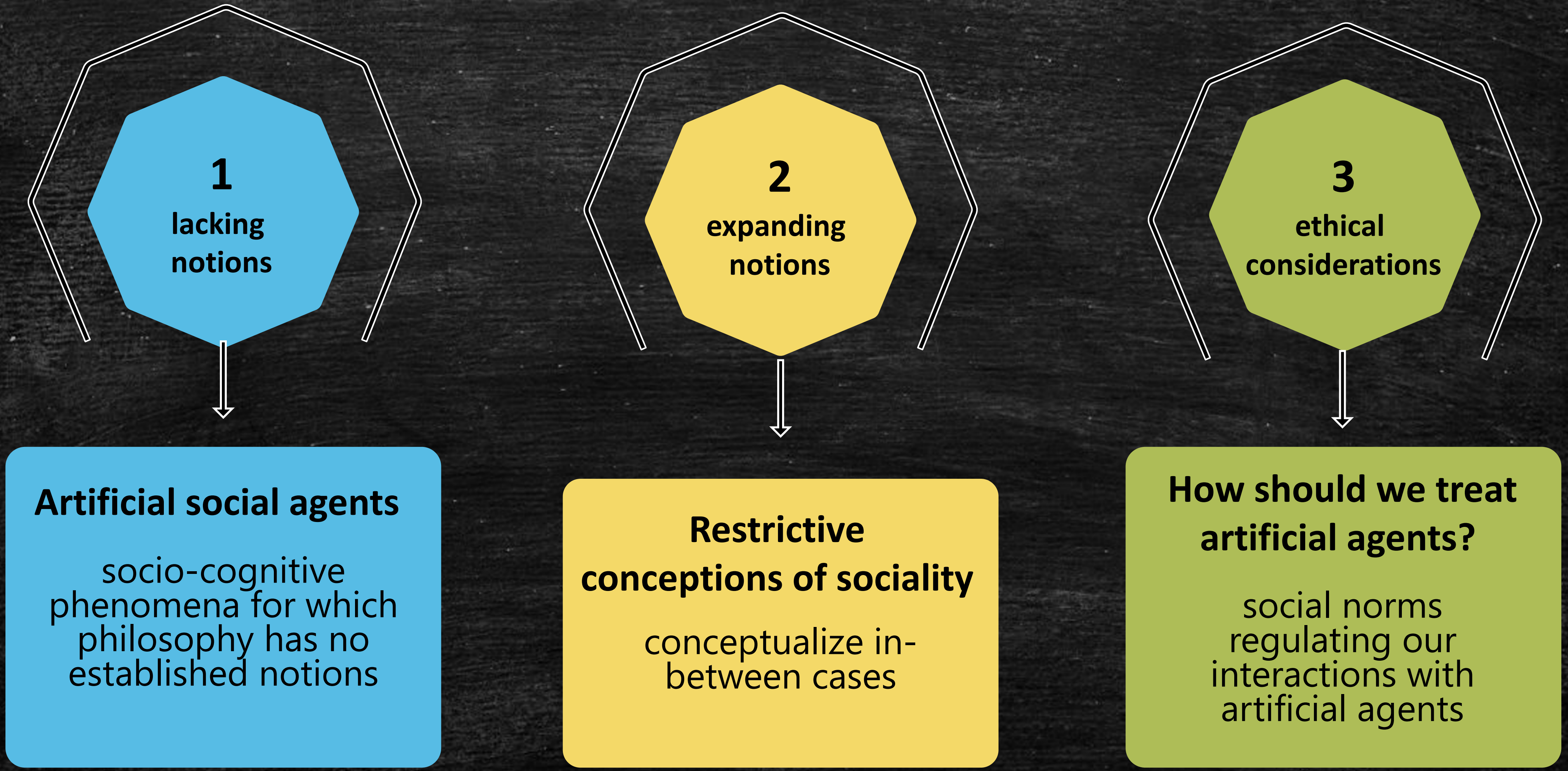
ARTIFICIAL AGENTS IN OUR SOCIAL WORLD

ANNA STRASSER

DENKWERKSTATT
BERLIN

OVERVIEW



1
lacking notions

Artificial social agents

socio-cognitive phenomena for which philosophy has no established notions

2
expanding notions

Restrictive conceptions of sociality

conceptualize in-between cases

3
ethical considerations

How should we treat artificial agents?

social norms regulating our interactions with artificial agents

1

tool-use

???

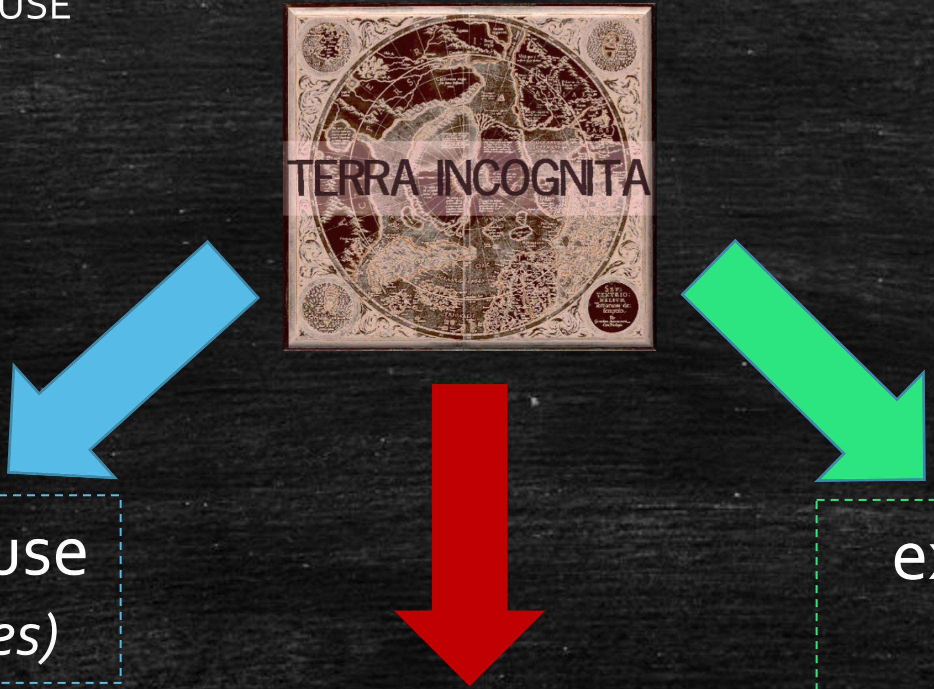
social interactions

1

No notions for in-between cases

NOT ALL HUMAN-MACHINE INTERACTIONS CAN BE REDUCED TO MERE TOOL-USE

SOCIAL INTERACTIONS SEEM TO REQUIRE LIVING AGENTS



expand concept of tool-use
(by integrating social features)

introduce a new category

expand conception of social interactions
(add non-living agents)

1

Reasons to expand the concept of social interactions

- similarity to human-human interactions
    - interactions with robots, avatars or chatbots
- humans connect emotionally to artificial agents
- used in empirical research to explore human behavior
    - experimental paradigms include human-machine interaction

→ investigate the limits of restrictive, standard notions & explore this terra incognita in order to conceptualize socio-cognitive phenomena with artificial systems

1

Joint actions

[Bratman 2014]

many demanding conditions

shared intentions & goals

specific belief state

relation of interdependence & mutual responsiveness

common knowledge

mastery of mental concepts

sophisticated mentalization skills

terra incognita:

- joint actions with non-human animals, infants and robots

ing Service, New Haven, Conn.

ONLY HUMANS NEED APPLY

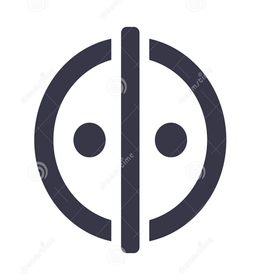
2

Restrictive conception of sociality

- lack of specific notions for in-between cases
- no theoretical grounds on which we can account for sociality of non-living beings

bound to Western conceptions

shintoism & animism



AND

notion of a social agent has proven to be changeable

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

2

How to overcome restrictive conceptions of sociality?

MINIMAL APPROACHES

→ MINIMAL VERSIONS OF STANDARD NOTIONS CAN CAPTURE A WIDER RANGE OF SOCIO-COGNITIVE ABILITIES

(1) assuming multiple realizations

→ questioning demanding conditions

- not all conditions necessary in the human case turn out to be necessary for artificial systems

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

- minimal mindreading (Butterfill & Apperly 2013)
- minimal sense of commitment (Michael et al. 2016)
- minimal action (Strasser 2006)
- shared intention light (Pacherie 2013)

2

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

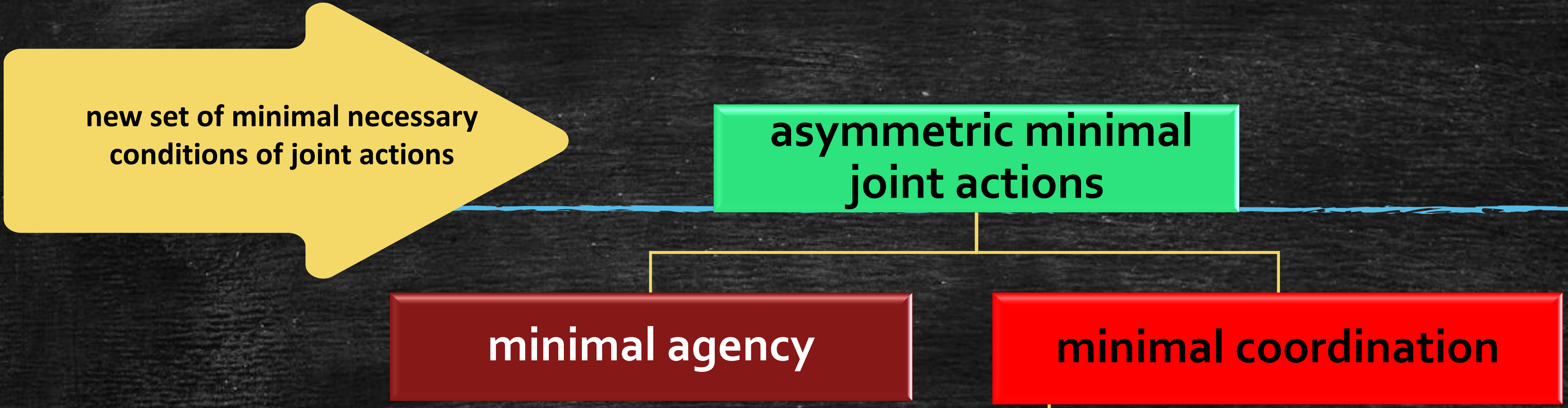
ARTIFICIAL INTELLIGENCE

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

ADULT & CHILD

ROBOT & HUMAN

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

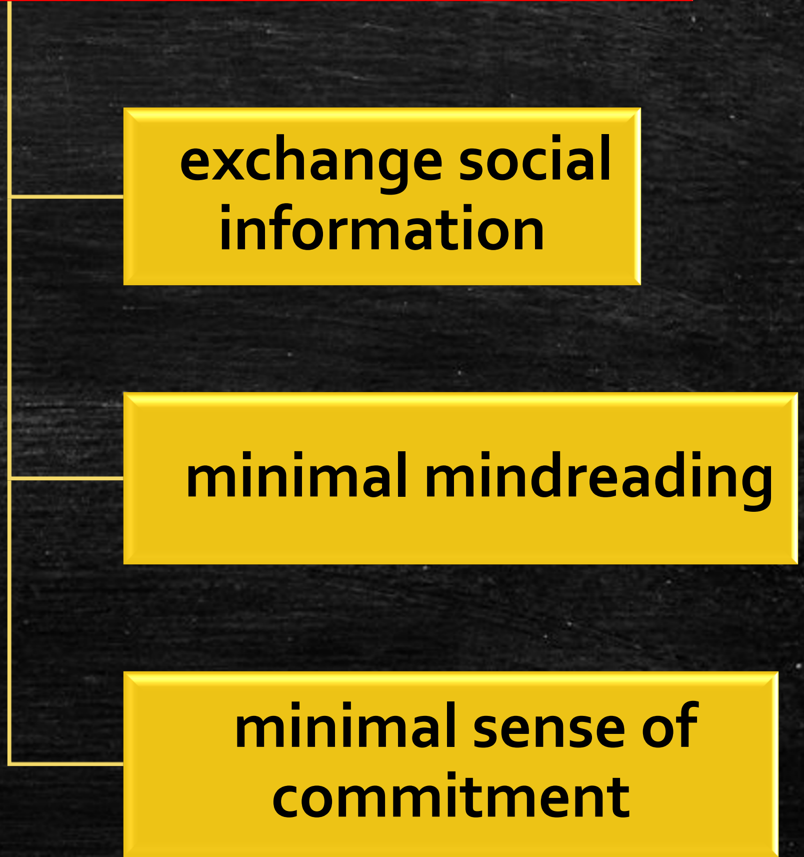


new set of minimal necessary conditions of joint actions

asymmetric minimal joint actions

minimal agency

minimal coordination



exchange social information

minimal mindreading

minimal sense of commitment

- assuming that biological constraints are not necessary for minimal agency

Strasser [2006, 2018]

Anna Strasser

Kognition künstlicher Systeme

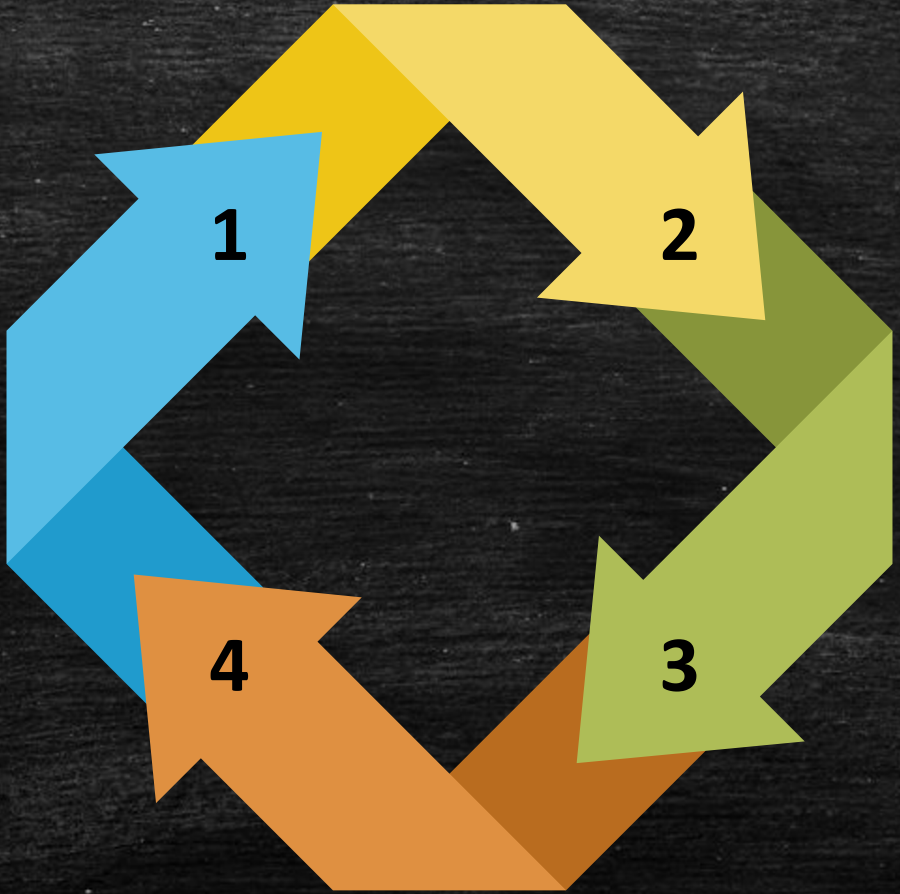
Philosophy & Cognitive Science

ontos verlag

exchange social information

← participants need social competences to coordinate

e.g., communication



humans send signals

verbal expression / social cues

systems process

interpret verbal expression / social cues presented by their interacting partners

systems respond

send verbal expression / social cues

humans process

interpret verbal expression / social cues presented by their interacting partners

affective loop [Höök 2009]

minimal mindreading

minimal necessary conditions for tracking others' perceptions & beliefs without representing perceptions & beliefs as such and without relying on conscious reasoning

ascribing less complex mental states

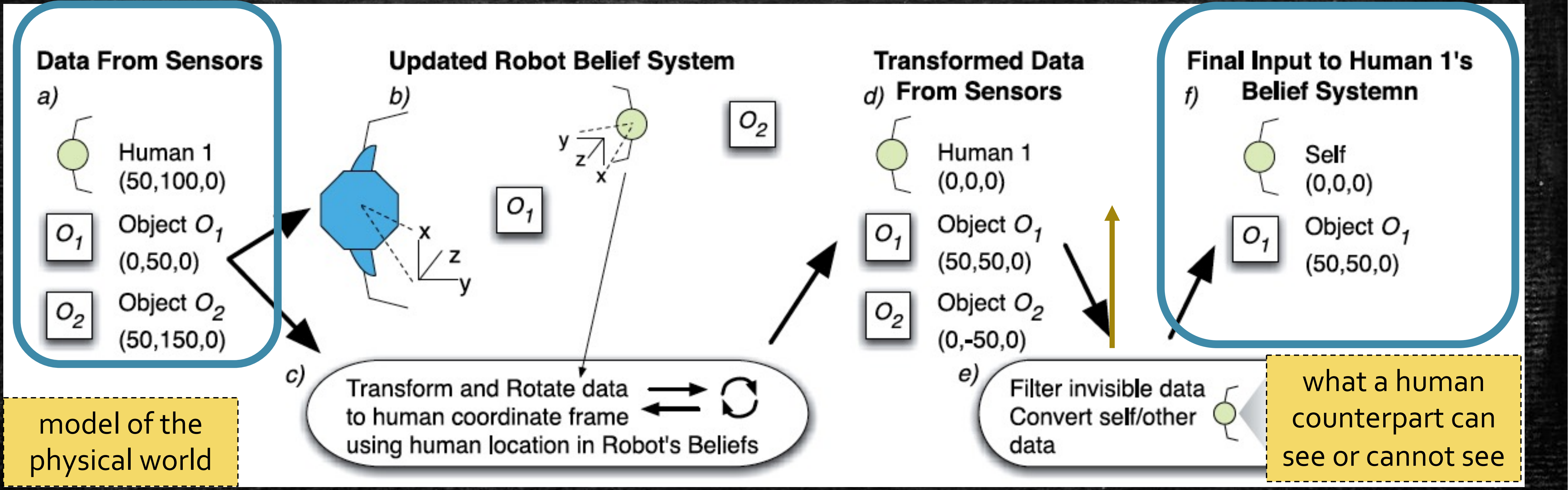
- encounterings *(kind of simple perception)*
- registrations *(rudimentary form of believing)*
- underlying operations : implicit, nonverbal, automatic, unconscious reasoning

Butterfill & Apperly [2013]

Minimal mindreading & artificial agents

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

Gray & Breazeal [2009, 2014]



Data From Sensors
a)
Human 1
(50,100,0)
Object O_1
(0,50,0)
Object O_2
(50,150,0)
O_1
O_2
Updated Robot Belief System
b)
x
z
y
O_1
O_2
c)
Transform and Rotate data to human coordinate frame using human location in Robot's Beliefs
Transformed Data From Sensors
d)
Human 1
(0,0,0)
Object O_1
(50,50,0)
Object O_2
(0,-50,0)
e)
Filter invisible data Convert self/other data
Final Input to Human 1's Belief Systemn
f)
Self
(0,0,0)
Object O_1
(50,50,0)

model of the physical world

what a human counterpart can see or cannot see

This human-centric representation will be used to anticipate future behavior of the human.

minimal sense of commitment

FOR MUCH OF WHAT COUNTS AS SOCIAL INTERACTIONS

- provides security humans need to rely on each other
- supports success of mindreading

Instead of requiring that an agent is only committed if she has assured her commitment and the other agent has acknowledged this, they claim that

- **components (expectation or motivation) of a standard commitment can be disassociated**

→ a single occurrence of just one component can be treated as a sufficient condition

Michael et al. [2016]

Consequences

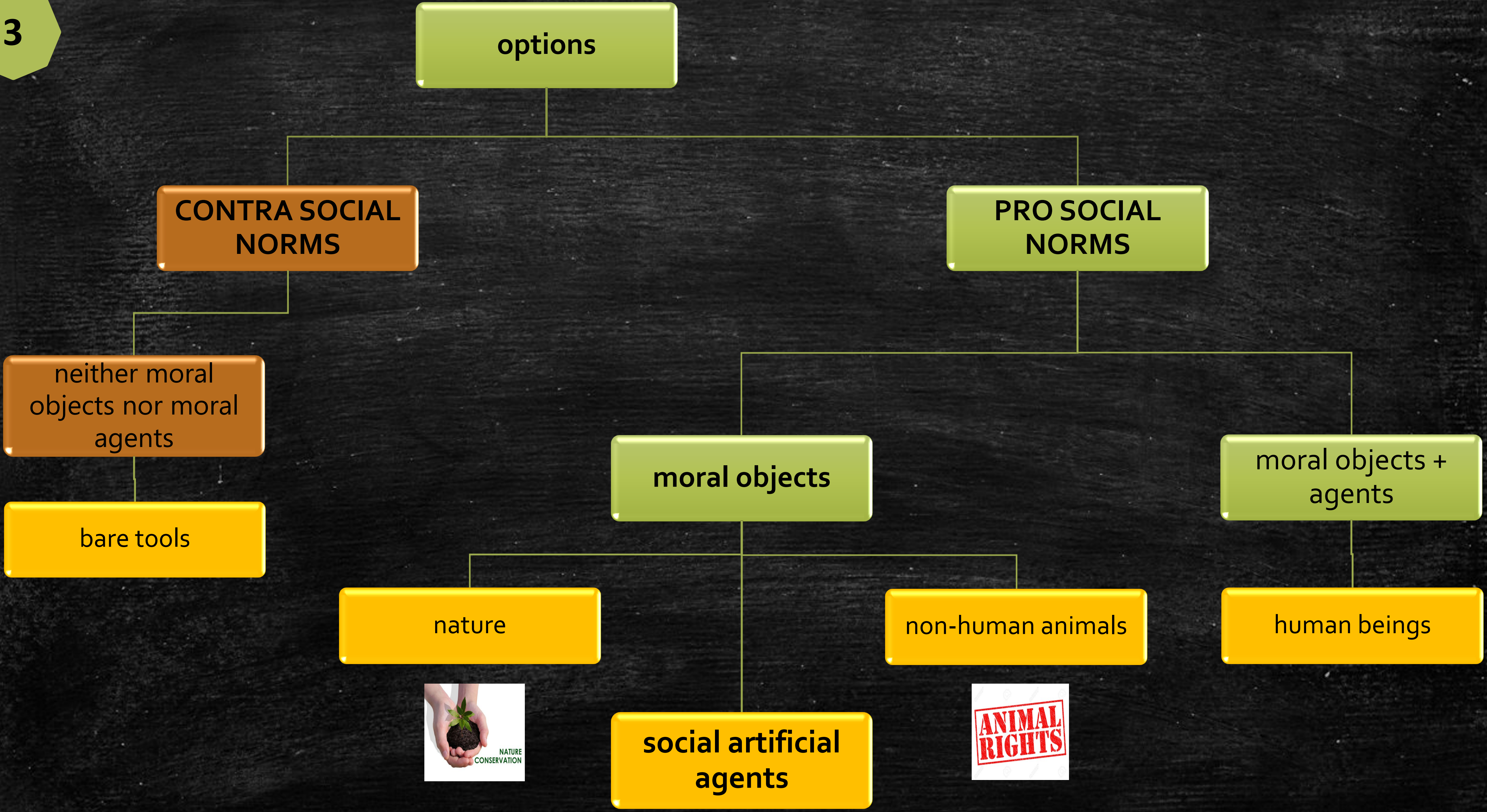
3

How should we treat artificial agents?

Social norms regulating our interactions with artificial agents

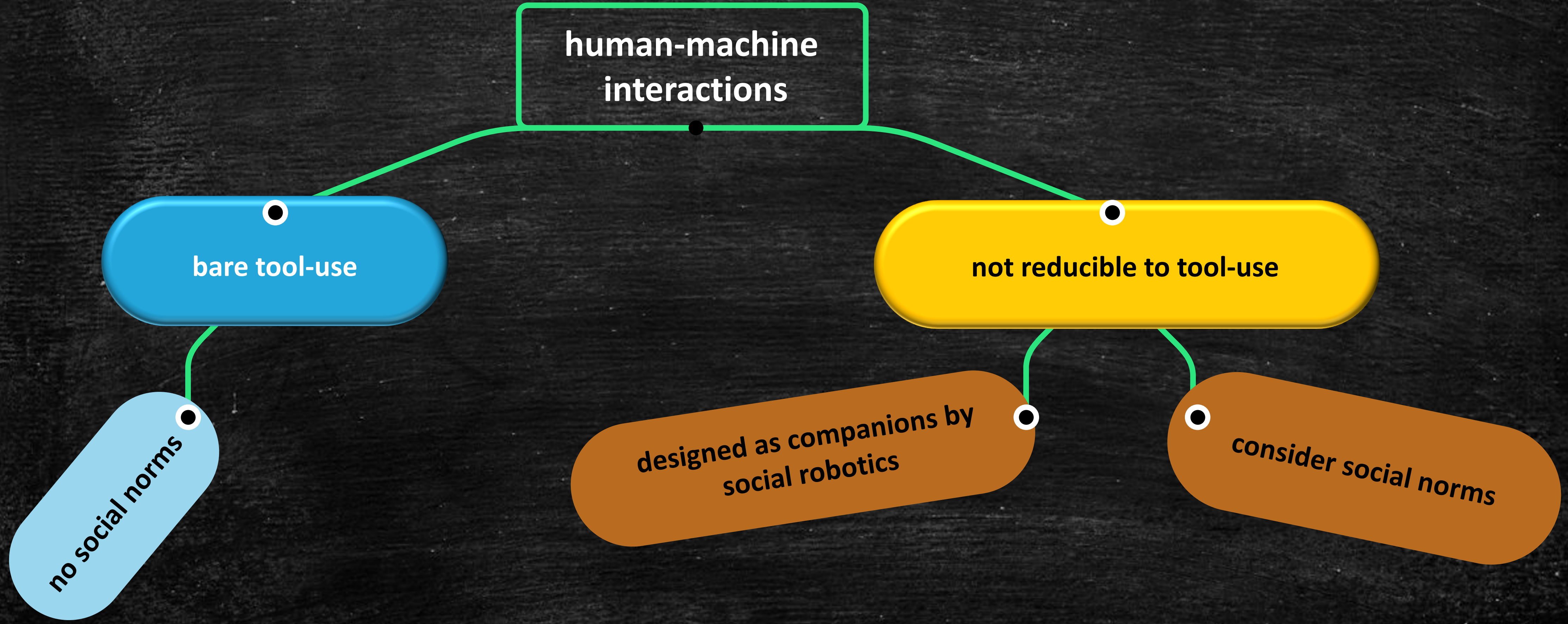
Is there something like morally appropriate behavior towards artificial systems?

3



- options
  - CONTRA SOCIAL NORMS
    - neither moral objects nor moral agents
      - bare tools
  - PRO SOCIAL NORMS
    - moral objects
      - nature
      - social artificial agents
      - non-human animals
    - moral objects + agents
      - human beings

3 Varieties of human-machine interactions


human-machine interactions
bare tool-use
not reducible to tool-use
no social norms
designed as companions by social robotics
consider social norms

We should be prepared that some human-machine interactions are not only part of our social life but also have the potential to change interpersonal interactions.

3

Arguing for social norms

IF THERE IS A PROBABILITY OF TRANSFERS

* behavior practiced with artificial agents can be transferred to other contexts

THEN VIOLATIONS OF SOCIAL NORMS IN-BETWEEN HUMANS ARE POSSIBLE

* through transferred behavior

THEREFORE, BEHAVIOR TOWARDS ARTIFICIAL AGENTS BECOMES A MORAL DIMENSION

NEW SOCIAL NORMS CAN PREVENT CRUEL MISBEHAVIOR AMONG HUMANS

social human-machine interactions

make human-machine interactions

morally relevant

for human-human interactions

3

1

APPLY BEHAVIORAL PATTERNS LEARNED IN A PARTICULAR CONTEXT TO A WIDE RANGE OF OTHER SITUATIONS

ABILITY OF GENERALIZATION

behavior practiced with artificial agents can be transferred to other contexts

AN ANECDOTES ABOUT ALEXA

- an older woman treated her ALEXA with particular politeness because she was afraid that she would lose her politeness towards humans as a consequence

3

2

PEOPLE TEND TO ACT ACCORDING TO SOCIAL NORMS WITH REGARD TO CERTAIN ARTIFICIAL SYSTEMS

ANTHROPOMORPHIZING

moral concerns regarding behavior toward certain artificial agents
make those interactions more similar to human-human interactions

We cannot help it but respond socially, even though our philosophical notions may tell us that our counterparts are not really social agents.

Kate Darling (2016):

- people are reluctant to behave destructively towards a toy robot(pleo)

3

3

PROPERTIES OF ARTIFICIAL SYSTEMS

SOCIAL ROBOTICS

Artificial systems contribute to a similarity with human-human interactions

- similarity to human-human interactions - a desired goal
- social robots should enter the human space of social interactions
- specific human-human interactions serve as models

To be successful as social interaction partners, research is being conducted to ensure that artificial systems have specific recognition systems, reasoning mechanisms, and the ability to initiate actions and also interpret social signals.

3

Desirable & disastrous transfers

positively evaluated

interactions in virtual reality

SKILL TRANSFER

- therapeutic settings using Virtual Reality for the treatment of phobias [Peperkorn et al. 2016]
- pilots transfer skills gained in a flight simulator

interactions with social robots

- people with dementia

negatively evaluated

interactions with sex-robots

- promote ubiquitous idea that even living women are sex objects → increase physical & sexual violence against women [Cox-George 2018]

interactions with paedobots

- lowers the inhibition threshold → makes it more likely to attack children

3

➡ A moral dimension

IF we notice that our behavior towards social robots can have a negative impact on our interpersonal behavior,

THEN we are motivated to think about whether we should regulate this behavior before it can be transferred.

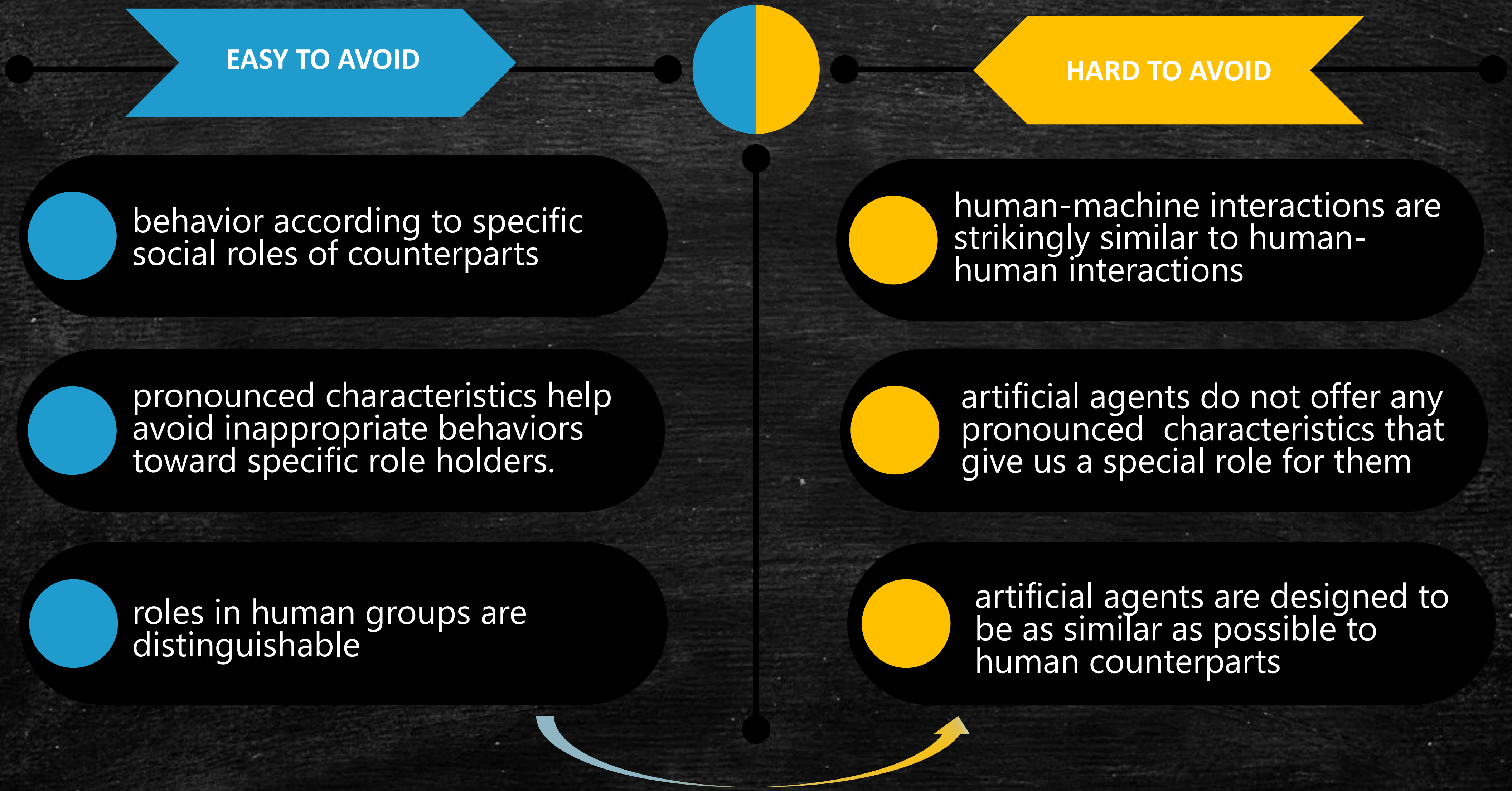
"One reason why humans might want to prevent the 'abuse' of robot companions is to protect social values."

Kate Darling (2016)

Even if the actual behavior towards an artificial agent is usually not morally evaluated, it can get a moral dimension by the possibility of a transfer.

3

Avoiding transfers



EASY TO AVOID

- behavior according to specific social roles of counterparts
- pronounced characteristics help avoid inappropriate behaviors toward specific role holders.
- roles in human groups are distinguishable

HARD TO AVOID

- human-machine interactions are strikingly similar to human-human interactions
- artificial agents do not offer any pronounced characteristics that give us a special role for them
- artificial agents are designed to be as similar as possible to human counterparts

POSSIBLE CONFUSIONS ARE ALMOST PREPROGRAMMED

CONCLUSION

IF you agree that destructive behavior towards nature cannot be classified as morally uncritical

- you may also agree with me that neither moral agency nor suffering capacities are necessary for counting as a moral object
- consequently, destructive behavior towards artificial systems cannot in principle be seen as morally uncritical

→ **Following a consequentialist strategy, one can claim that certain artificial systems can qualify as moral objects without capacity for suffering or moral agency.**

FUTURE RESEARCH

- Analyzing factors supporting the transfer of behavioral patterns
  → detailed assessment of the risks regarding transfers
- Investigation of how this risk might be reduced
- Evaluation of how this knowledge may shape our future construction of devices of social robotics

e.g., parents complain that their children unlearn
polite language like using ‘please’ and ‘thank you’

new Echo Dot Kids Edition: giving children positive
reinforcement when they say “please” and “thank you”

SUMMARY

CLAIM 1

- Some human-machine interactions are rather like social interactions than tool-use.

→ OVERCOMING RESTRICTIVE CONCEPTIONS OF SOCIALITY BY ESTABLISHING NEW NOTIONS.

CLAIM 2

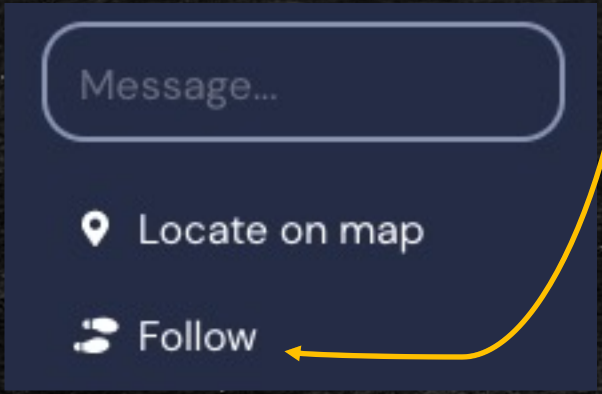
- specific (social) human-machine interactions can have an impact on human-human interactions

→WE SHOULD CONSIDER SOCIAL NORMS REGULATING OUR INTERACTIONS WITH ARTIFICIAL AGENTS!

Thanks a lot for your attention

Meet me in gather town

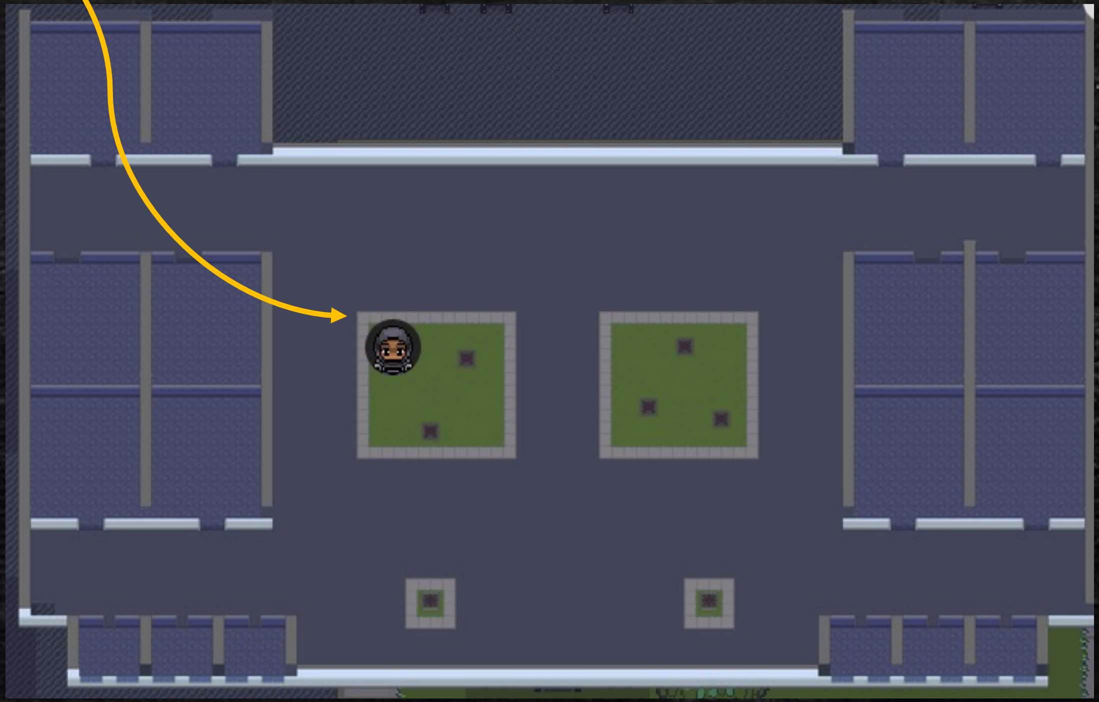
- click on participants icon
- find my name / click on it
  - *then you can locate me on the map or*
  - *automatically follow me*



Message...

Locate on map

Follow



references

- Baur, T. et al. (2015). Context-aware automated analysis and annotation of social human-agent interactions. *ACM Trans. Interact. Intell. Syst.*, 5, 2.

- Bratman, M. (2014). *Shared Agency: A Planning Theory of Acting Together.* Oxford: Oxford University Press.

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

- Carpenter, J. (2016). Culture and human–robot interaction in militarized spaces: a war story. London: Routledge.

- Cox-George Ch. & Bewley, S. (2018). I, Sex Robot: the health implications of the sex robot industry. *BMJ Sexual & Reproductive Health* 44, 161-164.

- Darling, K. (2016). Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior toward robotic objects. In Robot law, eds. R. Calo, A. M. Froomkin, and I. Kerr, 213–231. Northampton, MA: Edward Elgar.

- Davidson, D. (1980). *Essays on actions and events.* Oxford: Oxford University Press.

- Gray, J. & Breazeal, C. (2014). Manipulating Mental States Through Physical Action – A Self-as-Simulator Approach to Choosing Physical Actions Based on Mental State Outcomes. Intern*ational Journal of Social Robotics*, 6(3), 315–327.

- Höök, K. (2009). Affective loop experiences: designing for interactional embodiment. *Phil. Trans. R. Soc. B*, 364: 3585–3595.

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

- Mossbridge, Julia, and Edward Monroe. 2018. Team Hanson-Lia-SingularityNet: Deep-learning Assessment of Emotional Dynamics Predicts Self-Transcendent Feelings During Constrained Brief Interactions with Emotionally Responsive AI Embedded in Android Technology. *Unpublished XPrize Submission*.

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

- Peperkorn, H.M., Diemer, J., Alpers, G.W. & Mühlberger, A. (2016). Representation of patients' hand modulates fear reactions of patients with spider phobia in virtual reality. *Frontiers in Psychology, 7, 268.*

- Strasser, A. (2006). Kognition künstlicher Systeme. Frankfurt: Ontos-Verlag.

- Strasser, A. (2015). Can artificial systems be part of a collective action? In: Misselhorn, Catrin (Ed.). *Collective Agency and Cooperation in Natural and Artificial Systems. Explanation, Implementation and Simulation.* Series: Philosophical Studies Series, Vol. 122. Springer.

- Anna Strasser (2018). Social Cognition and Artificial Agents. In: Müller, V. (ed.) *Philosophy and Theory of Artificial Intelligence 2017. PT-AI 2017.* Studies in Applied Philosophy, Epistemology and Rational Ethics (SAPERE), 44, 106-114, Berlin, Springer.