PHILOSOPHY & THEORY OF ARTIFICIAL INTELLIGENCE

PT-AI 2021 - 4th Conference on "Philosophy and Theory of Artificial Intelligence"
Gothenburg, 27-28 September, 2021

Organisation
Vincent C. Müller, Technical University of Eindhoven (& University of Leeds & Alan Turing Institute)
Ivica Crnkovic & Gordana Dodig-Crnkovic, Chalmers & University of Gothenburg

DAY 1

Shannon Vallor (Edinburgh U) : "The Digital Basanos: AI and the Virtue and Violence of Truth-Telling"	
Alberto Termine & Alessandro Facchini (U Milano/IDSIA) Towards a Taxonomy of Pragmatic Opacity for the XAI Practitioner	Olle Häggström (Chalmers U) Artificial general intelligence and the common sense argument
Juan Duran (TU Delft) Trusting the output of black-box algorithms: A survery on computational reliabilism	Michael Cannon (TU Eindhoven) An Enactive Approach to Value Alignment in Artificial Intelligence
Tom Sterkenburg (LMU Munich) Undecidability in machine learning: What does it tell us?	Leonhard Kerkeling (Ruhr U Bochum) Matthew Liao's Approach of Ascribing Moral Status to AI Systems – Overview and Problems
Michael Levin (Tufts U): "Intelligence beyond the brain: basal cognition of life in diverse problem spaces inspiration for AI"	
Hajo Greif (TU Warsaw) Models, Algorithms, and the Subjects of Transparency	Fabio Tollon (U Bielefeld) Unpredictable Futures: Why, and How, we are Responsible for AI
Laura Crompton (U Vienna) The problem of AI influence	Lydia Farina (U Nottigham) Artificial Intelligence Systems, Responsibility and Agential Self-Awareness
Jiri Wiedermann & Jan van Leeuwen (CAS, Prague) Validating Non-trivial Semantic Properties of Robots	Andras Kornai (TU Budapest) Deception by default
Alice Helliwell (U Kent) The Ethics of AI-Generated Artworks	Guido Loehr (TU Eindhoven) Robot rights, grounded

Artificial general intelligence & the commonsense argument

2 REASONS REGARDING BEING RELAXED ABOUT THE WIENER-TURING WARNING

(a) AGI is unlikely or impossible in the foreseeable future.

(b) Surely a superintelligent AI would understand the wrongness of hurting us.

BUT THE RELEVANT QUESTION IS NOT

When AI exceed us in all cognitive domains?

→ RATHER →

*When will AI exceed us in **enough** domains to be better than us at **control** of the world?*

e.g., language natural processing is a substantial part …

AlphaZero is not a huge concern here, because finite two-player zero-sum board games with full information constitute perhaps... 0.1% of the range of important cognitive capacities?

On the other hand, text generation constitutes... more like 30%?

DANGER

If we cannot intervene anymore → we are at the mercy of the purposes of AI

We should be sure that the purpose of AI is really what we want!

An Enactive Approach to Value Alignment in Artificial Intelligence

Enactive Paradigm Primer: "Being defines a domain of relevance"

1. Autopoiesis
2. Cognition as "sense-making"

ENACTIVE VALUE ALIGNMENT

Claim: For highest "bandwidth" alignment, make AI as ontologically similar to humans as possible

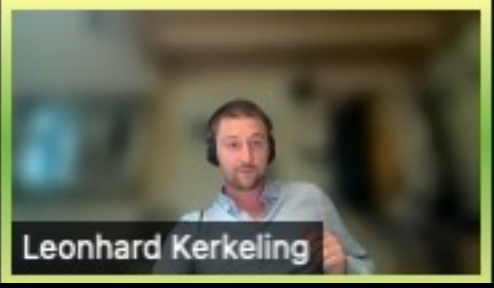
What defines what is relevant & not?

Existing Approaches	Enactive Approach
"Easy" Problem of Alignment	"Hard" Problem of Alignment
How we define the problem so that AI solves the right problem/doesn't solve the wrong problem?	**How do we make relevant to AI what is relevant to humans?**
• Problem-*solving* (computation) • Low/ "Thin" Bandwidth - aligned reasoning • Exogeneous value *specification* • Theoretical and empirical Comp sci, neuro sci, decision theory…	• Problem-*defining* (sense-making) • Relevance • High/ "Thick" Bandwidth – ontological alignment • Endogenous value *realisation* • Metaphysics & ontology, metatheory (integration of 1st & 3rd person epistemologies)…

VALUE ALIGNMENT:
How do we make relevant for AI what is relevant for humans?

ANSWER:
Make AI ontologically similar to humans

Matthew Liao's Approach of Ascribing Moral Status to AI Systems – Overview and Problems



Liao's approach solves *en passant* the question whether and when AI systems may befit moral status qua substrate neutrality: if an AI has the physical basis for moral agency it is due moral status.

Definition moral status: "[A]n entity has moral status when, in its own right and for its own sake, it can give us reason to do things such as not destroy it or help it." (Kamm 2007: 229)

→ Implies the necessity of some sort of intrinsic property held by an entity to have moral status.

Sentience, the capacity for phenomenal experience or qualia.

Sapience, set of capaacities associated with higher intelligence.

Moral agency, the capacity to act in light of moral reason.

physical basis

moral status

3 OBJECTIONS

1. **Circularity**
2. **Uninformativity/Limitation of Scope**
3. **Implausibility**

Unknown which genes make up the physical basis for moral agency we can hence not know what "functionally similar" for non-organic code actually means.

Argument: mere physical basis for moral agency seems in some cases not enough to justify a certain status; and with it trumping other's interests.

Intelligence beyond the brain:
basal cognition of life in diverse problem spaces inspiration for AI

cognition outside the brain

Biology exhibits intelligence at every level, operating as unconventional agents in many spaces besides the 3D world of "behavior".

unconventional intelligence

- Morphogenesis is an ancient proto-cognitive process with much to teach us about intelligence and the scaling of cognition. Life exhibits intelligent problem-solving and goal-directed activity in many spaces, at many scales.

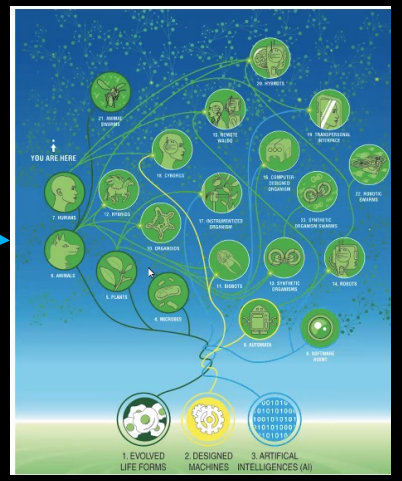
bioelectricity as cognitive binder

Somatic electrical activity is the cognitive medium of morphogenetic decision-making

- Bioelectrical networks underlie one type of up-scaling of cognition. This process shows us how multi-scale competency enables robust plasticity.

synthetic model systems

- dynamic, robust anatomical control = collective intelligence of cell swarms & we are now learning to read & write its medium



- Synthetic morphology and chimeric techniques
  - reveal plasticity and swarm cognition of cell groups
  - highlight our ignorance of large-scale outcomes given known subunits
  - provide a very rich new space for creation of novel, diverse minds which must be dealt with by theories of cognition
  - establish a (non-linear) continuum of bodies and minds which breaks current notions of machine, organism, robot, evolved/designed, etc. and requires us to identify the essential concepts, not those based on contingent evolutionary baggage (frozen accidents) and technological limitations of past eras

- Require new ethics for relating to novel forms of agency that aren't based on what the system is made of or its origin story

Unpredictable Futures: Why, and How, we are Responsible for AI

- Especially with machine learning systems, the correlations they uncover in data are *novel* (at least from the perspective of us cognitively limited human beings)
  - In the process of training these systems, engineers and programmers cannot predict the kinds of results that will be generated

Collingridge Dilemma

- the dilemma is that when a given technology is still in the nascent stages of development, it is possible to significantly influence the way it will develop, however, we lack knowledge of how the technology will affect society. Once the technology becomes "embedded" in society, and we come to know its implications, however, we are then in a position where we are unable to influence its development. In essence, when change is at its easiest, the need for it cannot be foreseen, and when change is required, it is difficult to implement (Collingridge 1980).

AI SYSTEMS DO NOT CREATE A UNIQUE GAP IN FORWARD-LOCKING RESPONSIBILITY

I supported this conclusion by focusing on the nature of risks when developing technology, and by showing that technological assessment is not only about the consequences that technology might have
This does not mean that forward-looking responsibility is not an issue when it comes to developing and deploying AI systems.

- Should be clear that AI does indeed complicate our responsibility ascriptions.
  - However, such complications do not lead to an insurmountable gap

Artificial Intelligence Systems, Responsibility and Agential Self-Awareness

AI responsibility is impossible (because they do not have consciousness)

The argument from consciousness

1. Responsibility requires/presupposes consciousness
2. AIs do not have a capability for consciousness

Therefore, AI responsibility is impossible

Rejecting premise 1: Responsibility requires/presupposes consciousness

Rejecting premise 2: AIs do have a capability for consciousness

Necessary Conditions for Responsibility:

- Responsibility requires:

Agential Self-Awareness: having an awareness of oneself as the one performing an action (Sebastian 2021)

self-representation →
having a self → minimal self
with or without consciousness?

- The minimal self: the subjective experience of having a self without entailing consciousness (Sartre 1957; De Beauvoir 1947; Husserl 1952).
- The minimal self entails **consciousness** (Garcia-Carpintero 2017; Recanati 2007).
- The minimal self is a result of self maintaining organisms where the ability of the system to represent itself is important for the overall maintenance of the system (Sebastian 2018).

Robot rights, grounded

THE WRONG KINDS OF RIGHTS ARE DISCUSSED

- VOTE
- NOT TO BE SOLD
- OWN PROPERTY
- A DIGNIFIED LIFE
- A FAIR TRIAL
- MARRY AND HAVE A FAMILY
- ASSEMBLE
- PRIVACY
- WORK
- ASYLUM

- **ABSURD:** ROBOTS HAVE THE RIGHT TO A FAIR TRIAL OR TO CHOOSE THEIR RELIGION?
- **UNREALISTIC:** ROBOTS WILL MOST LIKELY REMAIN UNCONSCIOUS
- **IMPRACTICAL:** WHO WANTS TO BUILD ROBOTS YOU CAN'T SELL?
- **CYNICAL:** MOST PEOPLE DO NOT HAVE THIS KIND OF PROTECTION

RIGHTS IN COOPERATIONS WITH ROBOTS

- Robot with an attitude
- Not sentient
- Owned by rental company
- Similar to idential behavior as human beings

RIGHTS/OBLIGATION TALK IS NOT MISAPPLIED IN THE CASE OF ROBOTS WITH ATTITUDE

- function of rights/obligation talk → makes conditions for cooperation conditions explicit
  → robots can make explicit what their conditions for cooperation are to us

DAY 2

Virginia Dignum (Umeå U): Responsible AI: from principles to action	
Oliver Buchholz (U Tübingen) A Means-End Account of Explainable Artificial Intelligence	Dan Weijers & Nick Munn (U Waikato) Human-AI Friendship: Rejecting the 'appropriate sentimentality' criterion
Gordana Dodig Crnkovic (Chalmers U) Cognitive Architectures Based on Natural Info-Computation	Elinor Clark (U Hannover) Decentring the discoverer: Rethinking agent-centred accounts of scientific discovery in light of advances in AI
Caterina Moruzzi (U Konstanz) Reaching Out-of-Distribution Generalization Through Robustness	Marcel Becker (Radboud U) Dignity in Digital Ethics
Kaisa Kärki (U Helsinki) Autonomy of attention	Carina Prunkl (U Oxford) Is there a trade-off between human autonomy and system autonomy?
Gualtiero Piccinini (U Missouri, St. Louis) Ontic Pancomputationalism and Computational Structuralism	Ralf Stapelfeldt (FU Hagen) Is it likely that you are living in a computer simulation?
Roman Yampolskiy (U Louisville) AI Risk Skepticism	
David Papineau (KCL, U London): "A Philosopher's Reactions to GPT-3"	

Responsible AI: from principles to action

EXAMPLES

manipulating, nudging chatbots, accountability of decision-making, power relations, dilemmas …

TRUSTWORTHY AI

Not *innovation vs regulation / ethics* but

Regulation/ ethics as stepping stone for innovation

Taking an ethical perspective
- Business differentiation ("Ethics is the new green")
- Certification to ensure public acceptance

Principles and regulation are drive for transformation
- Better solutions
- Return on Investment

REGULATIONS AND MORE

Regulation (EU)
- AI Act: Human-centred, risk-based approach

Standards (IEEE, ISO)
- soft governance; non mandatory to follow
- demonstrate due diligence and limit liability
- user-friendly integration between products

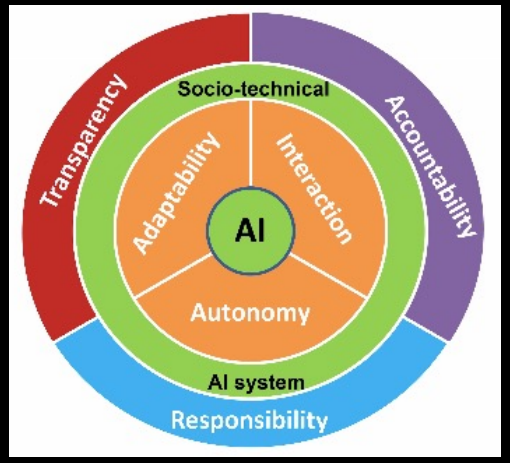
Advisory panels and ethics officers (Industry)
- Set and monitor ethical guidelines
- able to veto any projects or deliverables that do not adhere to guidelines

Assessment for trustworthy AI (EU)
- responsible AI is more than ticking boxes
- Means to assess maturity are needed

Awareness and Participation
- Education and training
- Appeal to civic duty / voluntary implementation

RESPONSIBLE AI IS LAWFUL, ETHICAL, RELIABLE, BENEFICIAL, VERIFIABLE

ART



CONCLUSIONS

- AI can do a lot! But use responsibly
- Deploying AI requires understanding what and why use it
- AI is not magic, but tools / artefacts made by people:

  We set the purpose
- AI can give answers, but we ask the questions

Human-AI Friendship: Rejecting the 'appropriate sentimentality' criterion

APPROPRIATE SENTIMENTALITY OBJECTION

1. Friendship requires appropriate sentimentality
2. AI cannot have the appropriate sentimentality
3. Therefore, AI cannot be friends

e.g. Helm (2017); Fröding & Peterson (2020)

QUESTIONING PREMISE 1

Strength/direction of sentiment doesn't necessarily correspond to the strength of a friendship

The value of caring sentiment is that it predicts and can cause caring intentions & behaviour

- But, caring sentiment doesn't always cause caring behaviour, it may even cause the opposite.
- Note: Replika Friends users say AI more reliably than human friends

Our account requires two features for friendship:

1. Mutual positive intentions
   - AI's can be programmed to include your wellbeing as a goal
2. A preponderance of rewarding interactions
   - People have this with AI, e.g., their AI-supported chat bots
   - AI can be programmed to recognize reward or receive manually

On our view, friendship is a concept of both kind and degree.

- Rejecting the appropriate sentimentality criterion for friendship, we argued that only mutual positive intention – the *attitude* of well-wishing is required to fulfil the non-experiential aspect of friendship.
- A consequence of this view is that if you find interacting with an AI rewarding and it wants good things for you, then it is a real friend.
- So, we don't need to worry about whether our new virtual friend is a human or really *feels* joy at our successes; it's enough that they continuously and sincerely do the things a friend should do because they wish us well.

Decentring the discoverer: Rethinking agent-centred accounts of scientific discovery in light of advances in AI

AC- accounts struggle to fit AI discoveries into this framework

CC- approach better able to deal with complexity of modern discoveries

AC-ACCOUNT CRITERIA

a) we can pick out a **relevant discovering agent**

b) who **conducted all**, or at least the important part, of the discovery process

BUT Locus of discovery unclear - hard to isolate one agent

c) and the discovering agent has **particular qualities/abilities** which are relevantly causally involved in the discovery

BUT AI lacks relevantly important abilities (Halina 2021; Stuart 2019)

COLLECTIVE VIEW OF DISCOVER

- there is no clear discovering agent who conducted all, or at least the important part, of the discovery process
- a collective of actors all made non-redundant contributions to the discovery
- credit for the discovery should be distributed between these agents depending on the contributions they made

- **Common knowledge** taken as given
- Just **current actors**
- **Distinction** between tools and non redundant, autonomous contributions

Is AI just a tool?

Then who is the discoverer? Creator? Interpreter?

Is there a trade-off between human autonomy and system autonomy?

Autonomous systems put human autonomy at risk *in virtue* of their increasing 'autonomy'

HUMAN AUTONOMY

The effective capacity of people to make decisions of their own that are of practical import to their lives.

- **Internal:** authentic values/decisions/motivations
  - No manipulation, addiction, …
- **External:** Freedom and opportunities
  - No coercion, compulsion, …
  - Existence of opportunities

SYSTEM AUTONOMY

- Independence from human operators (e.g. Franklin, 1996)

Examples: expert system, roombas, most airplane autopilots

- The ability to learn and act on the basis of experience (e.g. Russel and Norvig, 1998)

Examples: machine learning

Human Autonomy	System Autonomy
Internal dimension (authenticity)	Ability to learn
External dimension (freedom & opportunity)	Ability to operate independently

not undermining autonomy	undermining autonomy
delegation does not equal giving up autonomy • increasing opportunities *(autonomous wheelchair)* • increasing reflective capacities *(decision-*	online manipulation / deception / adaptive preference formation / surveillance

HUMAN AUTONOMY IN THE AI POLICY DISCOURSE

- **no unjustified coercion, deception, or manipulation**
(Ethics Guidelines for Trustworthy AI, High-Level Expert Group on Artificial Intelligence, 2019)

- **control over and knowledge about autonomous systems**
(Statement on Artificial Intelligence, Robotics, and 'Autonomous' Systems, European Group on Ethics in Science and New Technologies, 2018)

- **protecting human decision-making power**
(Floridi and Cowls, A Unified Framework of Five Principles for AI in Society, Harvard Data Science Review (1) 2019)

For there to be a trade-off, something needs to decrease as the result of something else increasing.

This is <u>not</u> the case for human autonomy and system autonomy.

Furthermore, the two concepts are fundamentally different in nature.

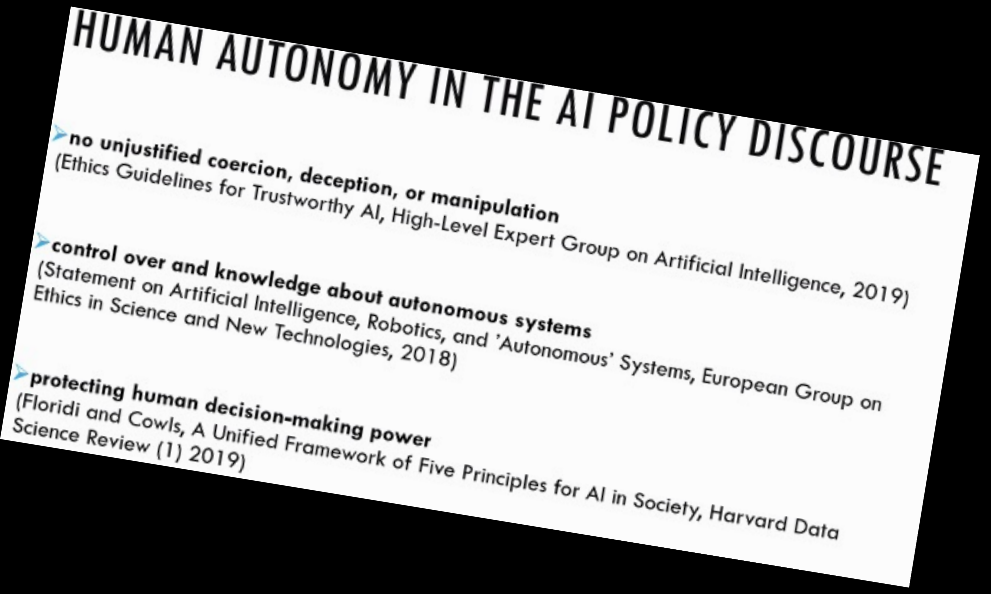
Protecting human autonomy in AI development nevertheless remains an important mission.

Is there a trade-off between human autonomy and system autonomy?

Autonomous systems put human autonomy at risk *in virtue* of their increasing 'autonomy'

HUMAN AUTONOMY	SYSTEM AUTONOMY
The effective capacity of people to make decision of their own that are of practical importance to their lives.	*independence from human operators*
INTERNAL (authentic values / decisions / motivations – no manipulation / addiction)	Ability to learn
EXTERNAL (freedom & opportunities (no coercion / compulsion / existence of opportunities)	Ability to operate independently

not undermining autonomy	undermining autonomy
delegation does not equal giving up autonomy • increasing opportunities *(autonomous wheelchair)* • increasing reflective capacities *(decision-making aids)*	• online manipulation • deception • adaptive preference formation • surveillance



For there to be a trade-off, something needs to decrease as the result of something else increasing.

This is <u>not</u> the case for human autonomy and system autonomy.

Furthermore, the two concepts are fundamentally different in nature.

Protecting human autonomy in AI development nevertheless remains an important mission.

Is it likely that you are living in a computer simulation?

BOSTROM' CLAIM

It is almost certain that we are living in a computer simulation!

1. Humanity will not be doomed, and will generate reallife ancestor simulations of human history.
2. Mental states are substrate independent.
3. Functionalism is true.
4. Computationalism is true.
5. Mankind will either perish or reach a posthuman state.
6. What is technologically possible in principle will be achieved in practice.
7. The creation of artificial conscious beings on computers is possible in principle and will be realized in a posthuman future.
8. There will be astronomical computing capacities.
9. The proportion of simulated persons among all persons ever existing is almost 100%.
10. The mathematical a priori argument is applicable.

IF one of the 10 background assumption does not hold

THEN it is almost likely that we do not live in a computer simulation

A Philosopher's Reactions to GPT-3

OTHER PHILOSOPHERS

CHOMSKY (1959) CONTRA SKINNER

Chomsky's review of Skinner's *Verbal Behavior* might contain compelling arguments that GPT-3 must be just a silly trick

FODOR (2010) CONTRA DARWIN

- Fodor claims that Darwin is not right with regarding the mechanism of evolution

GPT-3

- an associationist machine that grasps English syntax
- not less linguistically intelligent than many people we might meet in the pub – rather a lot smarter
- based on next word generation
- a reinforcement-learning system

REQUIRE COMPOSITIONALLY / PRODUCTIVITY / SYSTEMATICITY (Fodor & Pylyshyn)

- BUT unclear whether this must apply to the internally way of working

BUT

POVERTY OF STIMULUS DOES NOT APPLY

- maybe children are born with many of the connection weights that GPT-3 has to learn

SPEAKING WITH MEANING

- being embedded in a wider linguistic community → using words that have meaning

BEING IN THE WORLD

- restricted to speaking – only aim to guess the next word → BUT deep neural nets can have other aims (way finding / shopping…)

MODEL-BASED REASONING

- is achieved by representing causal structures in the world

→ Patrick Butlin (2021)

FORMAL EDUCATION

- competence in language can be the basis for further education

ETHICS & AI

- moral standing is not necessarily based on being conscious

HUMANS START RATHER DUMBS

- children and GPT-3 can be educated
- eventually even mathematics / critical thinking / history / engineering….