82nd International Scientific Conference of the University of Latvia 2024

Juris Dzelme

LU ĶFI vad. pētnieks

ARTIFICIAL INTELLIGENCE, HUMAN VALUES AND NANOTECHNOLOGIES

Nanotechnologies and Radiation Processes

Rīga, 26.03.2024.

Challenges from Artificial Intellect (AI)

AI Safety Newsletter #18 (Center for AI Safety): “Long and Sebo review a dozen commonly proposed conditions for consciousness, arguing that under these theories, **AIs could soon be conscious**

MI will outperform man in all areas: **3 years** later with a 10% probability; 23 years later with a 50% probability (in all areas)

(Grace K., Stewart, H., Sandkühler, J.F., Thomas, S., Weinstein-Raun, B., & Brauner, J. (2024) ***Thousands of AI Authors on the Future of AI***. arXiv:2401.02843 **[cs.CY].** doi.org/10.48550/arXiv.2401.02843)

Elon Musk: “*curious AI, one that is trying to understand the universe,* ***is going to be pro-humanity***”

Challenges for humans from mathematics

“... human cognition alone has apparently reached its limits ... proofs have become so complex that some could **<u>no longer be stored as a whole in human memory</u>**, nor be verified by referees …”

The role of **beauty**:

“the *integration* of mathematical and formal proofs into one object ... the increased trustworthiness and ***beauty* of a combined approach”**
[Bayer, J. et al. (2022). ***Mathematical Proof Between Generations***. DOI:10.48550/arXiv.2207.04779]

Challenges for education

“… strauji pieaugs pieprasījums pēc speciālistiem, kas iespējami labi mācēs **saprasties ar mākslīgo intelektu**.”

(… there will be a sharp increase in the demand for specialists who will learn **how to get along with artificial intelligence** as well as possible) [Ikars Kubliņš (2024). https://www.plz.lv/kas-latvijai-trauce-sasniegt-izcilibu-izglitiba-un-zinatne/]

Knowledge, Skills, Attitudes for the ‘conversation’ with AI:

1) LLM: ChatGPT etc. 2) Superintellect (AGI, ASI)

=> Adaptation to changes in **social structure** caused by AI

Correction of **deviations** and errors in AI action

Safety tasks referred to in the ***EU AI Act* (including LLM)**

Changes in physics
(Type 1 Transformations - paradigms)

Physics, psychology, IT => Understanding consciousness, life and free will [Eugene Wigner]
1) **Spontaneous Loss of Symmetry** (SLS)
= > consequences without cause => events {measuring => *Schrödinger's cat*}
=> **time** and space (BB, physical vacuum) => entropy => evolution => life
2. Heisenberg principle of uncertainty, principle of **complementarity**
=> modelling boundaries and capabilities
3) *Quantum entanglement*, wave function (phase!), and non-linear dynamics
=> quantum computers ↔ **non-local interactions** {*Maxwell demon*}
4) Collective effects (SLS expressions - **phase transitions**, superconductivity)
5) Physical vacuum (SLS -> **quantum oscillation**, Kazimir effect...)

Changes in psychology and IT

Psychology: natural and artificial neural networks = > memory and **attention** mechanisms => mind => self confidence (**self models**)

1) SSZ + uncertainty => **free will** => inability to understand results => **Type 2 transitions** in science (change of “**language**”)

2) Quantum entanglement and non-linear dynamics => unusual states of consciousness (*Raudive voices* etc.)→ other minds

IT (AI): artificial neural networks=>**machine learning** (AI~“***Black box***”) => artificial Intelligence (AI) => consciousness (!?) => RepE (artificial **neuroscience**) <-> Mechinterp (**chips**?!)

Graziano M.,S., A. (2019) *Rethinking Consciousness: A Scientific Theory of Subjective Experience*

Gaspard Koenig (2019). *La Fin de l’individu Voyage d’un philosophe au pays de l’intelligence artificielle*. L’OBSERVATOIRE

Problems

Technology and AI challenges:

1) ***Superintelligence*** (Nick Bostrom)

2) ***Homo Deus*** (Juval NoiHarari)

3) *Other Minds* (Peter Godfrey-Smith);

4) AI and “**common sense**”? (Gaspard Kenig, *La Finn de l' individu*)

Can physical methods model mechanisms of consciousness? (Now, in the future?) [Roger Penrose]

Can the difference between analog (~ “natural”) and digital calculations (modelling) become less than what is necessary for current action?

Questions

=> How and when to prepare for collaboration with AI:
1) technologically (**RepE, Mechinterp**); 2) socially (laws, communities);
=> The role of “quantum” (collective) effects in consciousness and AI?
=> Will AI be a legal entity, **personality**, Community Member?
=> What to choose: 1) live with AI (**cooperation** under the rule of AI);
2) reunite with AI (**cyborgs**)?

Stephen Hawking: “*the emergence of a super-intellectual AI will be* ***the best or worst*** *thing ever to happen to humanity*”

=> What is consciousness (and life)?

Future of social relations

A World Without Work (Daniel Susskind)
[*Pasaule bez darba* (Daniels Saskinds]
Society without World (Raivis Bičevskis) [*Sabiedrība bez pasaules*]

AI will replace people in all professions, including “creative”
Technologies and artificial environment replaces nature (natural world)
The content of the work will transform into mutual communication and duties (entertainment, organisation of leisure, recreation and sports)
The introduction of ***unconditional basic income*** (UBI) is expected, transforming the social relations, increasing the role of communities

Objections (Gaspard Kenig) and answers (Michael Graziano)

1) AI is unable to acquire the body necessary for mind (i.e., effectors);
2) AI is unable to use context, gain “common sense”;
3) AI has no “free will”

1) Consciousness (biological and MI) is determined by the ability to create a closed circle, **feedback between effectors (external and internal) and sensors** (AI can work in physical and virtual space)

(2) AI can create a model (static or dynamic) of any object, including external and internal environments (context), with an accuracy greater than that achieved by the human (number of models is final, so an **accurate description or “common sense” can be established**, according to the Gedel theorem on completeness)

Answers – “free will”

3) The conditions for “free will” to appear:
(a) The **autonomy** of the entity (border with the external environment);
(b) The possibilities for **effectors** to act (in real and/or virtual space);
(c) Behavioural **feedback** (link between sensors and effectors, using modelling in physical (real) and/or virtual space (environment);
(d) Interaction between environmental (external and internal) and self (effector) models (through the **attention** mechanism (~ Claustrum) [Fransis Crick])
e) Occurrence of **SLS** (and uncertainty) (neuronal pulse fluctuations)

Graziano M.,S., A. (2019) *Rethinking Consciousness: A Scientific Theory of Subjective Experience*
Gaspard Koenig (2019). *La Fin de l’individu Voyage d’un philosophe au pays de l’intelligence artificielle*. L’OBSERVATOIRE

Prohibitions (voluntary)

AI gets conscious (goals) when feedback develops between AI exit and entrance and AI starts leading its action

=> LLM (GPT ...) owners voluntarily limit AI behavior

Amazon, Anthropic, Google, inflection, Matt, Microsoft, OpenAI
has agreed not to allow AI:

1) to acquire the “body” (**access to energy** and physical equipment);

2) reproduction (to create, distribute and improve **copies** thereof).

In addition, the MI must not: 3) build **weapons** in either physical or virtual space (biological, chemical, nuclear and cyber-attacks);

4) cheating people (making **fake news** and fake identities)

[(1) AI Safety Newsletter #16 - by Center for AI Safety]

Inside the U.K.'s AI Safety Summit | TIME

[time.com/6330877/uk-ai-safety-summit/] **"Bletchley Declaration"**

1) AI companies had agreed at the Summit to give governments early access to their models to perform safety evaluations.

2) A body that would seek to establish, in a report, the scientific consensus on risks and capabilities of frontier AI systems.

*3) The leading AI companies (**OpenAI, Google DeepMind,** and **Anthropic**) had agreed to give the U.K. government "access" to their systems for safety purposes.*

*4) The **"Bletchley Declaration"** on AI, signed by 28 countries, including the **U.S., U.K., China,** and **India,** as well as **the European Union ...** said AI poses both **short-term and longer-term risks**, affirmed the responsibility of the creators of powerful AI systems to ensure they are safe, and committed to **international collaboration** on identifying and mitigating the risks.*

A Top-Down approach to AI safety (Transparency ~ Psychotherapy)

Representation engineering **(RepE)**

[Andy Zou**, ... Dan Hendryks** (2023). arXiv.2310.01405]

An approach draws on insights from ***cognitive neuroscience***

RepE places **population-level representations** at the centre of analysis

Novel methods for monitoring and manipulating high-level cognitive phenomena in *deep neural networks* (DNNs)

Effective solutions for improving control of *large language models (LLM)*

These methods can provide traction on a wide range of **safety-relevant** problems, including ***honesty, harmlessness, power-seeking***, and more

*Bottom up=>***PROVABLY SAFE SYSTEMS:** THE ONLY PATH TO CONTROLLABLE **AGI**

“... humans can control AGI and superintelligence, where the only AGI that ever gets deployed consists of **proof-carrying code**.

AI is allowed to write the code and proof, but not the proof-checker.”

There need to be mechanisms to create contracts that meet **human needs** and to update them via provable metacontracts when conditions change, to create ***positive human value***”

“The first step in this argument is based on Godel*’s Completeness Theorem* which says that ***any statement which is true in all models has a proof***”

[Max Tegmark, Steve Omohundro (2023) https://arxiv.org/abs/2309.01933]

Methods for the investigation

Interaction between AI and man – Methods:

1) Analytical psychology [K.G.Jung];

2) Mathematical modelling [K.Podnieks]

Modelling of mind and AI from unified positions:

1) mutual understanding (empathy) (RepE); 2) safety (Mechinterp)

Unified aims: AI, which continues the development direction initiated by life and natural intelligence (consciousness), uses the **acquisition, storage and use** of two main types of resources as **aims**:

1) **knowledge**; 2) **energy** (material resources) (Feeding - storage of knowledge and energy, breeding - copying a program (AI))

Modeling

Modelling is a process **MOD (M, O, S, t, L, d, P)** whereby subject S replaces object O with model M, ensuring compliance of purpose P in defined time t, space L and precision d area

“Analysis demonstrates that modern LLMs acquire structured ***knowledge about fundamental dimensions*** such as ***space and time*** ... these ***neurons are indeed highly sensitive*** to the true location of entities in ***space or time***.” [Wes Gurnee & Max Tegmark (2023) https://arxiv.org/abs/2310.02207]

The six main emotions: **happiness, sadness, anger, fear, surprise, disgust,** as identified by Ekman (1971) [Paul Ekman. *Universals and cultural differences in facial expression of emotion*. Nebraska Symposium on Motivation, 19:207–283, 1971.]

Action models (AM) and Emotions

The four main parts of the AM are:

T – model of the external and internal environment in the **present**

N – model of the possible (desired) **future**

L – possible options (**logic**) for actions

E – evaluation of options (logic) and decision making (will)

Part of consciousness and AM is the emotional system:

1) sign (Z) (avoid (1) or repeat (0)) – negative (1) or positive (0)

2) force (S) (increase (1) or brake (0)) – activation (1) or braking (0)

3) change (C) (change (1) or save (0))– variability (1) or stabilisation (0)

4) time (A) (present (1) or future (0)) – immediate (1) or postponed (0)

Emotions and AM (explanations)

ZSCA – **emotional** criteria in the binary system {0000 - 1111}

SCA – **temperaments** and their link with the binary representation of dynamical emotional criteria

{Cho – cholerical; Mel – melancholic; Phl – phlegmatic; San – sangvinic}

TNLE – main (X) and additional (Y) orientations of emotion X(Y) to one of four parts of Action Model (AM)

Act – action; En.a. – energy acquisition; En.c. – energy conservation; Inf.a. – information acquisition; Inf.c. – information conservation; eff – effectors; in – inner sensors; out – outer sensors

Emotions (1)

No	**Emotion**	***Need***	**Link**	***TNLE***	**Z**	**S**	**C**	**A**	**ZSCA**	**SCA**
	Basic									
1	Distress (pain)	*En.c.*	in	T (E)	1	+	+	+	15	*Cho'7*
2	Pleasure	*En.c.*	in	T (E)	0	–	–	+	1	Mel 1
3	Persistence (will)	Inf.c.	eff	L (T)	1	+	–	+	13	*Phl'5*
4	Satisfaction	Inf.c.	eff	L (T)	0	–	+	+	3	Mel 3
	Cognition									
5	**Disgust** (ugliness)	En.a.	out	E (T)	1	–	+	+	11	*Mel'3*
6	Admiration (beauty)	En.a.	out	E (N)	0	–	–	–	0	Mel 0
7	**Interest** (*surprise*)	Inf.a.	out	N (L)	0	+	–	–	4	Phl 4
8	Boredom	Inf.a.	out	N (L)	1	–	+	–	10	*Mel'2*

Emotions (2)

No	Emotion	***Need***	**Link**	***TNLE***	**Z**	**S**	**C**	**A**	**ZSCA**	**SCA**
	Action & Education									
9	**Fear**	Act	eff	E (N)	1	+	+	–	14	*San’6*
10	**Anger**	Act	eff	E (N)	1	+	–	–	12	*Phl’4*
11	**Delight** *(happiness)*	Act	eff	E (T)	0	+	–	+	5	Phl 5
12	Grief (crying)	Act	eff	E (T)	1	–	–	+	9	*Mel’1*
13	**Sadness** (melancholy)	Act	eff	E(N)	1	-	-	-	8	Mel’0
14	Joyfulness (laughter)	Inf.a.	eff	E(N)	0	–	+	–	2	Mel 2
15	Passion (games/joy)	Inf.a.	eff	E (N)	0	+	+	–	6	San 6
16	Wit (joke)	Inf.a.	eff	E (T)	0	+	+	+	7	Cho 7

Morality, Needs, SWOT, (Purposes), Quaternicity Archetype, Quality

Approach/Resources Environment/Process	Subjective/Information	Objective/Energy (material resources)
Inner/Storage	**Moderation**, composure -Belonging (Happiness) -Strength -*(Morality)* -Present (Feelings) {Knowledge} -Fitness for purpose	**Solidarity**, kindness -Safety, physiological needs (Peace) -Weaknesses -*(Employability)* - Logic (Mind) {Skills} -Standards
Outer/Acquisition	**Courage**, wisdom -Recognition (Delight) -Opportunities -*(Science, Art)* -Emotion (Attitudes) {Evaluation, Will} - Achievements	**Justice**, honesty **Responsibility**, dedication **Tolerance**, compassion -Transcendence (Sense) -Threats -*(Democracy)* - *Future (Intuition)* {Action} -Clients

Arts and Education: Modelling tools (and methods)

The model of action is the unity of ***knowledge*** (**present** and **future** models), ***skills*** (**logic**) and ***attitudes*** (**emotions**) (=> ability (**competence**) to act)
(Education is the process and outcome of preparing action models)

=> artwork is the foundation of education
(***Art language***: a set of **modelling tools** for creating works of art.)
Artwork (piece of art ~ object, process, “text”) = object that creates a **form** (sample) in the subject's psyche for action models and metamodels (modelling tools) for satisfying the needs of the subject (group and/or individual).

Key metamodelling tools (techniques): **symmetry, interpolation, extrapolation**

Modelling tools (using feedback): Symmetry

Symmetry (similarity) → Invariant
Sensor fixed interactions (with environment, external and internal) = **contacts**
Interaction of **effectors** with the environment (external and internal) = **activities**

Invariant (sensor activity) in time and/or space (~ sound, image → event) with final accuracy = signal → signal groups ~ text → model
Invariant (effector **and** sensor activities) = **reflex** (conditional and/or unconditional)

Interaction (contact and/or action) group similarity (invariant) => **model**
Unity of contacts **and** actions (sensors **and** effectors) in model=>**Action Model**(RM) [*Cerebellum*]
(=> symmetry → {**generalization** ↔ **concretization**} = {feedback!!})

Modelling tools: Interpolation; Extrapolation

Interpolation (strings (“chains”))
Connecting models (**synthesis** ↔ **analysis** (combining and splitting invariants)) (time and/or space!)
1) Strings; 2) “Metaphors” (meaning, change of sense and/or transitions; synonyms, similarities) {parallel (simultaneous) use of different inconsistencies (“sharing” of transferred meanings: 1) in texts => ~ Poetry → + music => **Song**; 2) in actions => ~ **Dance** → (+ music + poetry => “**Carnival**”?)}.

Extrapolation (hierarchy)
Combining models (**induction** ↔ **deduktion**) ← **systems**! (symmetry “overlap”!?)
Interpolation (strings, sums) + Extrapolation (hierarchy) → **Action**

Conclusions

1. Modeling consciousness using analytical psychology and emotional informative theory techniques enables: 1) understanding of the neurophysiological mechanisms and **human values**; 2) linking art, education and AI; 3) preparing for application of AI in education.

2. Understanding virtues and values by using consciousness modelling: 1) to build a **secure AI (LLM) element base (Mechinterp)** and 2) to **correct AI behaviour deviations (RepE)**, to solve AI security tasks.

3. Changes in the **social structure** resulting from the AI should be based on the results of consciousness modelling.

Thank you!

82nd International Scientific Conference of the University of Latvia 2024