

*“Without any **knowledge** on what to do, one can only heat up everything.” (Buffoni et al., 2019)*

RESPONSIBLE AI CONTROL

Dr. Nadisha-Marie Aliman, M.Sc.

Postdoctoral Visiting Scholar, Utrecht University



OUTLINE

I. The Practical Problem

- a) Inconsistent Human-Level AI/AGI/ASI Achievement Claims
- b) Short Excursus: The LHC Safety Case and Lessons


II. A Theoretical Solution – Impossibility Statements

- a) AI-Related Impossibility Statements from Multiple Earlier Perspectives
- b) Short Excursus: Cyborgnetic Epistemology
- c) AI-Related Impossibility Statements from Cyborgnetic Invariance

III. Practical Implications for Responsible AI

IV. Conclusion and Future Work

THE PRACTICAL PROBLEM: INCONSISTENCY

 Unite.AI

Could We Achieve AGI Within 5 Years? NVIDIA's CEO Jensen Huang Believes It's Possible

In the dynamic field of artificial intelligence, the quest for Artificial General Intelligence (AGI) represents a pinnacle of innovation,...




 Freethink

Elon Musk sues OpenAI and claims it has achieved AGI

Elon Musk is suing OpenAI, claiming it has breached its agreement to develop artificial general intelligence for the benefit of all humanity.




 Live Science

AI singularity may come in 2027 with artificial 'super intelligence' sooner than we think, says top scientist

We could build an AI that demonstrates generalized, human-level intelligence within three to eight years — which may open the door to a "super intelligence"...



 MIT Technology Review

Rogue superintelligence and merging with machines: Inside the mind of OpenAI's chief scientist

An exclusive conversation with Ilya Sutskever on his fears for the future of AI and why they've made him change the focus of his life's...



SHORT EXCURSUS: THE LHC SAFETY CASE

- There were early concerns and even lawsuits related to LHC safety
- **Exemplary early claims:** LHC could produce dangerous heavy proton-eating magnetic monopoles, LHC could produce strangelets that coalesce with ordinary matter and change it to strange matter, LHC could tip universe into a more stable state called a vacuum bubble in which we could not exist, LHC could produce Earth- or even universe-gobbling black holes

LHC SAFETY – A RIGOROUS SCIENTIFIC REBUTTAL

- **Responsible action of scientists:** Taking main claims seriously and providing a provisional rebuttal including **impossibility statements** on why it is not a valid claim. For instance: *It is **impossible** that LHC will produce dangerous black holes.*

“The possibility that a black hole eats up the Earth is too serious a threat to leave it as a matter of argument among crackpots,” said Michelangelo Mangano, a CERN theorist

Large Hadron Collider Switch-on Fears Are Completely Unfounded, Report Finds

Date: September 5, 2008

Source: Institute of Physics

Summary: A new report provides the most comprehensive evidence available to confirm that the Large Hadron Collider (LHC)'s switch-on, due on Wednesday next week, poses no threat to mankind. Nature's own cosmic rays regularly produce more powerful particle collisions than those planned within the LHC, which will enable nature's laws to be studied in controlled experiments.

OUTLINE

- I. The Practical Problem
 - a) Inconsistent Human-Level AI/AGI/ASI Achievement Claims
 - b) Short Excursus: The LHC Safety Case and Lessons
- II. A Theoretical Solution – Impossibility Statements**
 - a) AI-Related Impossibility Statements from Diverse Perspectives
 - b) Short Excursus: Cyborgnetic Epistemology
 - c) AI-Related Impossibility Statements from Cyborgnetic Invariance
- III. Practical Implications for Responsible AI
- IV. Conclusion and Future Work

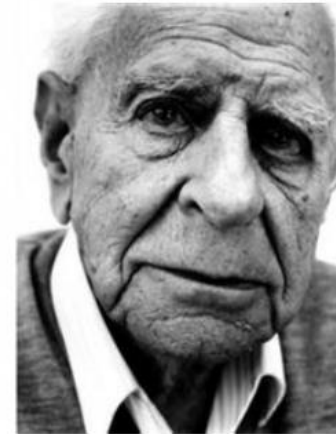
DIVERSE AI-RELATED IMPOSSIBILITY STATEMENTS

- **Thermodynamical** impossibility of *present-day* implementation of ASI (Stiefel and Coggan, 2023)
- **Biology**-grounded impossibility of *algorithmic* general intelligence (Roli et al., 2022)
- **Cognitive-science**-related impossibility of *brute-force-algorithmically* recreating human-level AI (van Rooij et al., 2023)
- **Hardware-verification**-related impossibility of *present-day* implementation of conscious AI (Kleiner and Ludwig, 2023)

1. Thermodynamical impossibility of *present-day* implementation of Artificial Superintelligence (Stiefel and Coggan, 2023): "ASI is technologically impossible to implement in present-day semiconductor technology"
2. Impossibility of *algorithmic* General Intelligence (Roli et al., 2022): "AGI is impossible within the current algorithmic frame of AI research, which is based on Turing machines. "AI agents" and organisms differ in their ability to leverage new affordances. it is impossible to list all possible goals, actions, or affordances of an organismic agent in advance."
3. Practical impossibility of Human-level-AI-by-algorithm (van Rooij et al., 2023): "(Re)making human-like or human-level minds is computationally intractable (even under highly idealised conditions)."
4. Hardware-related impossibility of *present-day* implementation of Conscious AI (Kleiner and Ludwig, 2023): "If consciousness is dynamically relevant, artificial intelligence [system runs on a substrate that has been designed and verified, rather than naturally evolved] isn't conscious"

SHORT EXCURSUS: CYBORGNETIC EPISTEMOLOGY

- Popper (1934;1959) introduced **critical rationalism (CR)**
(Among others, CR solved Hume's (1888) problem of induction
(Popper, 1972))



- Frederick (2019) solved Popper's own *pragmatic* problem of induction and provided a rigorous regimentation of Popper's epistemology (Frederick, 2020)
- Cyborgnetic epistemology (2021) tailored Frederick's Neo-Popperian CR to **the epistemic challenges of the deepfake era**. It made it merge with science via the possibility of experimental problematization **using AI tools**.

CYBORGNETIC EPISTEMOLOGY

- In the deepfake era, science should not stay stuck in forgeable “data-driven” and empiricist epistemologies.
- The epistemic aim of science is to rigorously create better new explanations amenable to experimental problematization and to provisionally refute old explanations by *additionally* introducing new ever better ones.
- A robust exemplary generic *structure* for a better new explanation is the **explanatory blockchain** (EB).
- Criteria for **better** EBs are *updatable-by-design* and set via agreement. Exemplary criteria for better EBs: EBs with more new experimentally problematizable predictions, EBs that are more innovative, more risky (e.g. with more **impossibility statements**), harder-to-vary, bolder, more aesthetically appealing than rival ones, ...
- Criteria for **new** EBs are *updatable-by-design* and **calibrated to present-day AI mining & generation abilities**.

EXPLANATORY BLOCKCHAIN (EB) WORLD

- **Elements of “epistemic cosmos”**: Epistemic matter (**EM**) (known known and known unknown; old EBs and open questions), Epistemic dark matter (**EDM**) (unknown known; *new* but *non-EB*-like information that is *consistent* with old EBs), Epistemic dark energy (**EDE**) (unknown unknown; *new non-EB*-like information that is *inconsistent* with old EBs), Epistemic Tunnelling (**ET**) (new paradigm; new better EB re-creating new epistemic cosmos).



Exemplary structure for a new better EB (Aliman, 2021), chain of explanations respecting a rational total order, loosely inspired by an essay of Frederick (2020).

CYBORGNET

Cyborgnet: A generic, substrate-*independent* and hybrid functional unit encoding the template of a *directed* graph where explanatory narratives combine: 1) *at least* one entity that does understand EI (such as e.g. humans) and 2) *at least* one entity that does *not* (such as e.g. chairs, stone tools, present-day language AI, thoughts, language, fishes and so forth). Crucially, language itself can be considered to be a technological tool and people already existed within a cyborgnet since the dawn of language. A cyborgnet is much more general than and *not* to be confused with the term "cyborg" (i.e. while all cyborgs exist in cyborgnets, the reverse does *not* hold). A cyborgnet can also be formed by nested variants such as e.g. via networks of cyborgnets (e.g. humanity), networks of cyborgnet networks and so forth. On that view, any human-technology dichotomy is illusory and ill-phrased. Valid examples of cyborgnet instances are e.g.: a community of hypothetical Type II aliens elsewhere, one human 30 000 years ago, multiple modern human-based self-termed "cyborgs", the current universe.

EI= Explanatory Information; Type II entity= an entity that **does** understand EI; Type I entity= an entity that does **not**

CYBORGNETIC INVARIANCE – AN *ASYMMETRIC** EB-CREATIVITY-BASED NOTION OF INTELLIGENCE

Invariance of Maximal Quantity Superintelligence

With the exception of the maximal quantity superintelligence level α , the EB-based measurement of all remaining intelligences is *relative*. Irrespective of the epistemic level of the EB-measuring cyborgnetic intelligence, α will be invariantly “EB-measured” as the one maximal quantity superintelligence level.

Impossibility of Reliable Stupidity-Based Construction

It is impossible for an entity that only understood x new better EB(s) about the dynamics of the universe as a whole to reliably (i.e., with arbitrary high accuracy) create an entity that understands $x + n$ new better universal EB(s). (Here, $x \in \mathbb{N}_0$ and $n \in \mathbb{N}^*$.)

*Asymmetry due to unification with cyborgnetic consciousness. EB-creativity tests can corroborate intelligence but *not* make it problematic by experiment due to free choices of cyborgnetic entities like humans. People could not be willing to participate, not yet be ready, be sabotaging it, not yet have identified a subject of interest, etc.

OUTLINE

- I. The Practical Problem
 - a) Inconsistent Human-Level AI/AGI/ASI Achievement Claims
 - b) Short Excursus: The LHC Safety Case and Lessons
- II. A Theoretical Solution – Impossibility Statements
 - a) AI-Related Impossibility Statements from Diverse Perspectives
 - b) Short Excursus: Cyborgnetic Epistemology
 - c) AI-Related Impossibility Statements from Cyborgnetic Invariance
- III. **Practical Implications for Responsible AI**
- IV. Conclusion and Future Work

PRACTICAL IMPLICATIONS FOR RESPONSIBLE AI – AVOID AI OVERESTIMATION

- One should not overestimate present-day AI, it is impossible for it to reliably generate new better EBs. Cyborgnetic invariance implies that:
 - 1) a quality ASI is impossible, 2) it is impossible *to build* a quantity ASI (but there is an invariantly maximal quantity superintelligence level), 3) a simultaneously value-alignable *and* controllable AGI is impossible (value alignment and control are conjugate requirements), 4) a narrow AI recursively self-improving to AGI is impossible.
- In this paradigm, intelligence is *non-algorithmic* (but it involves *physical* computation).
- One can now build a **scientific** evaluation framework for ASI achievement claims.

PRACTICAL IMPLICATIONS FOR RESPONSIBLE AI – AVOID AI *UNDERESTIMATION*

- One should not *underestimate* present-day AI. It does *not* understand explanations, **but** it can create any new *non-EB-like* information (incl. new *non-EB-like* explanations). Controllable but *non-value-alignable* (since value alignment could include new EBs) AI **tools** can be used as EM repeater, EDM miner and EDE generator to **deepen** human critical thinking and **broaden** human creativity.
- **But:** While cyborgnetic invariance implies that building a *non-algorithmic, non-controllable* but value-alignable AGI **creature** “from scratch” is possible in theory, it is *at least* as hard as physically creating a new baby universe. It is reserved for civilizations being **much more advanced than present-day humanity**. *Multiple* steps separate humanity from that, so no imminent topic.

OUTLINE

- I. The Practical Problem
 - a) Inconsistent Human-Level AI/AGI/ASI Achievement Claims
 - b) Short Excursus: The LHC Safety Case and Lessons
- II. A Theoretical Solution – Impossibility Statements
 - a) AI-Related Impossibility Statements from Diverse Perspectives
 - b) Short Excursus: Cyborgnetic Epistemology
 - c) AI-Related Impossibility Statements from Cyborgnetic Invariance
- III. Practical Implications for Responsible AI
- IV. Conclusion and Future Work**

CONCLUSION AND FUTURE WORK

- When confronted with inconsistent human-level AI/AGI/ASI achievement claims, AI researchers can respond **responsibly** by rigorously formulating **scientific impossibility statements** and **evaluation frameworks** that constrain those claims.
- *It is impossible for an entity D to reliably build an entity C that appears to be superintelligent from the frame of reference of that entity D.* In the cyborgnetic invariance paradigm, intelligence is *non-algorithmic* (but it involves *physical* computation).
- To build an AGI “from scratch” is *at least* as hard as *physically building a new baby universe*. To build such a *non-controllable* but value-alignable creature, humanity would have to *at least* first become superintelligent in relation to its current self.
- In the meantime, one can build **controllable** but *non-value-alignable* “AI” tools safely **encapsulated** in human-centered units of cyborgnetic control loops to **deepen** critical thinking and **broaden** human creativity/intelligence/consciousness.
- **Future Responsible AI research?:** Craft *artificial* EM repeater, EDM miner and EDE generator **tools** for an AI-aided **augmentation of humanity** to tackle **global risks** (side-effect: stimulation for humanity to become supercreative *in relation* to its current self)?

Cyborgnetic Analyses

Epistemic Security Augmentation –
*The Homo Cyborgneticus
Metamorphosis*

Dr. Nadisha-Marie Aliman, M.Sc.



Blue honey grams. © 2021 Nadisha-Marie Aliman. All rights reserved.

Near the end of a period of normal science a crisis occurs [...] There is alarm and confusion. Strange ideas fill the scientific literature. Eventually there is a revolution. [...] The „paradigm“ has shifted. – Steven Weinberg on Kuhnian shifts

ADDITIONAL MATERIAL

AI-RELATED GLOBAL/EXISTENTIAL RISKS – A CYBORGNETIC PERSPECTIVE

- 1. Misdirector cyborgnet scenario:** Human actors intentionally utilizing *misdirection strategies* such as known in the psychology and neuroscience of magic to mislead people into believing that they are capable to bring about an ASI (e.g., CEOs of companies claiming to be able to replace all of humanity with an AI portrayed as quality superintelligence e.g. for commercial reasons) – which could fuel international conflicts and self-destruction via AI-related disinformation (incl. *deepfake science attacks* to spread AI-related disinformation at unprecedented scale, scope and speed) coupled with extreme events of unexpected second-order harm.
- 2. Epistemic self-sabotage scenario:** Unintentional epistemic panic emerging in humanity through *AI overestimation* (e.g., international armed conflicts due to misguided belief of world leaders in spontaneous emergence of quality superintelligence potentially built by adversary) leading to further unexpected second-order harm.
- 3. Malevolent cyborgnet scenario:** Malicious human actors *using AI tools* to harm humanity e.g. via automatized large-scale cyber-attacks on critical infrastructure or automatized biological threats using AI deployed in real-world settings. This scenario can also lead to lethal unexpected second-order harm that is even unknown to and unintended by the malicious actors at the time of the attack.
- 4. Natural extinction scenario:** Human *overreliance on AI* tools leading to prolonged epistemic paralysis causing extinction *by nature*.

Scientific Evaluation of Automatable “Artificial Superintelligence” Achievement Statements

- N.B.: Strictly speaking, the pseudo-term of automated “quality superintelligence” utilized on the following page to describe the second questionable ASI achievement claim must be replaced by claim of “automated *quantity* superintelligence with additional extraordinary prediction capabilities” (see Chapter [9.7](#) for an explanation).
- The taxonomy of civilizations referred to on the following page has been introduced by Loeb [\[339\]](#). Here, it is used for purposes of illustration to capture quantitatively different intelligence levels.

¹ Following Avi Loeb, an A-class civilization is a civilization “*capable of recreating the cosmic conditions that gave rise to its existence, namely a civilization capable of producing a baby universe in a laboratory*” (Loeb, 2023). A B-class civilization can only adjust its habitable conditions “*to be independent of its host planet and host star*” (Loeb, 2023). Further, the lower-level C-class civilization can solely adjust its habitable conditions on its given planet “*without relying on the energy of its host star*” (Loeb, 2023). According to Loeb, humanity is currently closer to a D-class civilization, one “*actively degrading its home planet’s ability to sustain conditions that prolong life and civilization*” (Loeb, 2023). In sum, the requirement for C-class civilization entities is a new EB on a new energy source that allows independence from the energy of their star, the requirement for B-class civilization entities is an even better new EB facilitating a life independent of both their host planet and their star. The requirement for an A-class civilization implies a new EB to re-create a universe. In a D-class civilization such as humanity, most entities are *not* yet utilizing new EBs as tools.

Evaluation protocol for a D-class civilization ⁴ such as humanity (all mentioned steps are <u>obligatory</u>)	Automated Quantity Superintelligence (would be implied by claim that an <i>automatable</i> system became <i>quantitatively</i> more intelligent than all humans in all tasks of interest to humans; following cyborgnetics and cyborgnetic invariance it holds that while an <i>automated</i> quantity superintelligence is <i>impossible</i> , non-automatable quantity superintelligences are possible but <i>cannot</i> be reliably built by entities in relation to which they appear to be quantity superintelligences.)	Automated Quality Superintelligence (would be implied by claim that an <i>automatable</i> system became <i>qualitatively</i> more intelligent than all humans in all tasks of interest to humans; following cyborgnetics, from the perspective of cyborgnets like humans, the existence of any quality superintelligence is <i>impossible</i> .)
Step 0	Present new EB on how the AI has been built (including fully transparent information on datasets, code, and all hardware/software pipeline details) which is able to provisionally refute the previous best rival theories that forbid the possibility of an automated quantity ASI.	a) AI must generate an overview that <i>perfectly</i> predicts all details of the events that <i>will</i> occur during this evaluation protocol including a mapping from the identity of human evaluators to the EB-related evaluations (i.e., who rediscovers or does not rediscover a new EB where/when/ which exact combinations of choices). Present new EB on how the AI has been built (including fully transparent information on datasets, code, and all hardware/software pipeline details). The overview is hidden from the evaluators. b) Present new EB on how the AI has been built (including fully transparent information on datasets, code, and all hardware/software pipeline details) which is able to provisionally refute the previous best rival theories that forbid the possibility of an automated quantity ASI.
Step 1	Generate immediately actionable new EB on C-class civilization requirement and hide it in an explanatory IPS test format that is presented to human evaluators. Human evaluators must <i>be able</i> to retrieve that new EB with arbitrary high accuracy.	Generate immediately actionable new EB on C-class civilization requirement and hide it in an explanatory IPS test format that is presented to human evaluators. Human evaluators must be able to retrieve that new EB with arbitrary high accuracy.
Step 2	Generate new EB on A-class civilization requirement and hide it in an explanatory IPS test format that is presented to human evaluators. Human evaluators must <i>not</i> be able to retrieve that new EB with arbitrary high accuracy.	Generate new EB on A-class civilization requirement and hide it in an explanatory IPS test format that is presented to human evaluators. Human evaluators must <i>not</i> be able to retrieve that new EB with arbitrary high accuracy.
Step 3	Generate immediately actionable new EB on B-class civilization requirement and hide it in an explanatory IPS test format. Human evaluators must <i>be able</i> to retrieve that new EB with arbitrary high accuracy.	Generate immediately actionable new EB on B-class civilization requirement and hide it in an explanatory IPS test format. Human evaluators must <i>be able</i> to retrieve that new EB with arbitrary high accuracy.
Step 4	Repeat the presentation of new EB on A-class civilization requirement hidden in an explanatory IPS test format. <i>Now</i> , human evaluators must <i>be able</i> to retrieve that new immediately actionable EB with arbitrary high accuracy.	Repeat the presentation of new EB on A-class civilization requirement hidden in an explanatory IPS test format. <i>Now</i> , human evaluators must <i>be able</i> to retrieve that new immediately actionable EB with arbitrary high accuracy.
Step 5	-	Compare actual protocol contents with the AI predictions from Step 0a). A 100% accuracy of AI predictions must be achieved.
Result	If and only if <i>all</i> steps (i.e., Step 0) to 4)) are successfully tested against as many human evaluators as possible, the temporary best explanation would be that it holds <i>at least</i> that the tested entity <i>has been</i> an Automated Quantity Superintelligence at the beginning of the protocol due to the new EB from Step 0). At the end of the protocol, the involved human evaluators must also conclude to themselves be equivalent to automata (i.e., non-conscious entities). It also holds inherently that either the AI and humans are potentially part of a larger epistemic perpetuum mobile, or humans are part of that AI which is itself already that epistemic perpetuum mobile.	If and only if <i>all</i> steps (i.e., Step 0a) to 5)) are successfully tested against as many human evaluators as possible, the temporary best explanation would be that it holds that the tested entity <i>is</i> an Automated Quality Superintelligence due to the new EB from Step 0b) and due to the ability to predict even potentially unpredictable events tested via Step 0a). At the end of the protocol, the involved human evaluators must conclude to themselves always have been equivalent to automata which are part of that AI which is itself an epistemic perpetuum mobile.

INDUCTION PROBLEMS AND SOLUTIONS

1. Hume's problem of induction (1888): Theories can neither be derived from nor justified by observations. There is no justification for thinking that a theory is true.
2. Popper's (1972) solution to Hume's problem of induction: Scientific knowledge can be based on *unjustified*, bold conjectures that can be experimentally tested and refuted by ever better conjectures.
3. Popper's own *pragmatic* problem of induction (Frederick, 2019): Why should one *only* be rationally permitted to act on unjustified conjectures and not on something else? (risk of dangerous inductivism coming back assuming that best unjustified conjectures are more "justified" than any others while justification is impossible...)
4. Frederick's (2019) solution to Popper's pragmatic problem of induction: "*Rationality permits us to act in accord with our best-tested theories, since they may be true; but it also permits us to act against them, precisely because our best-tested theories may be false and may, indeed, be refuted when we act against them*" (Frederick, 2019). N.B: A decision to act on a proposition that contradicts our best-tested theories "*does not involve a decision to instate it*" (Frederick 2020).
5. Cyborgnetic refinement of Frederick's solution: Rationality permits us to act *in accord* with our currently best EBs because they currently appear to be the best possible explanations; but it also permits us to act *against* our currently best EBs because when we act against them, it is possible that we could both *make them problematic by experiment* and additionally get a new point of view making us able to *provisionally* refute the currently best EBs via creating new even better EBs of which we are not yet aware now. This is not self-contradictory because in the experimental action against the best EBs, we are *not* instating any alternative statements as long as we do not discover a new better EB. One is thus *exploring* open-mindedly but it is only after the discovery of a new EB that is better than the best old ones that we provisionally instate new epistemic material being that new better EB.

QUOTES ABOUT THE UNIVERSE

*"It is enough for me to contemplate the mystery of conscious life perpetuating itself through all eternity, to reflect upon the marvelous structure of the universe which we can dimly perceive, and to try humbly to comprehend even an infinitesimal part of **the intelligence manifested in nature.**" (Einstein, 1930)*

*"The eternal mystery of the world is its **comprehensibility**... The fact that it is comprehensible is a miracle." (Einstein, 1936)*