Generic Analyses for Information Theory and Epistemology

Deepfake Science – Crafting Epistemic Defenses

Dr. Nadisha-Marie Aliman, M.Sc.

Deepfake Science – Crafting Epistemic Defenses

***Deepfake Wetenschap* –**

Epistemische Verdedigingen Ontwerpen

GENERIEKE ANALYSES VOOR

INFORMATIETHEORIE EN EPISTEMOLOGIE

Utrecht, Netherlands

Dedicated to the potential of cyborgneticity.

Contents

Chapter 1

Introduction

This book written as ephemeral mental clipboard has been assembled within a dense period of a few days. Here, *explanatory blockchain* [7] (EB) decryption cannot be sharply distinguished from EB encryption endeavors. For a better epistemic preparedness and in order to be able to map the terminology utilized in this book to pre-existing concepts, it thus seems recommendable to be equipped with the body of knowledge conveyed in at least the first three cyborgnetic books [7, 8, 5]. However, a familiarity with the first one may already improve a grip on many relevant key ideas. In this book, I provide new bold conjectures that pertain specifically to contemporary *cybersecurity-oriented AI safety* and *epistemic security* in the *deepfake era* and more generically to *epistemology* and *information theory.* The modus operandi differs from previous ones in that it explicitly immerses all performed harm analyses in a *cyborgnetic retrofuturistic* [5] setting framed in a counterfactual multiverse called *Cyland.* This involves both a process of projecting a counterfactual future as seen from the past and a process of projecting a counterfactual past as seen from the future. I utilize design fiction narratives from a fictive city called *Cynam* located in the Cyland multiverse, to answer the following major questions:

1. Do the information-theoretical limitations of present-day Type I AI risk to harm the knowledge creation processes of humans (being Type II entities[1])?
2. How could humans design Type I AI that could instead be used to mitigate epistemic threats such as *deepfake science* attacks in a more systematic manner?
3. Why could science profit from a currency based on Type II cynetbits and embedded in a so-called Type II *cynetbit*coin blockchain?

[1] Type II entities are all entities able to *understand* explanatory information. The only Type-II-*species* on Earth is humanity. (For rare cases with *individuals* from other species see [7].) There may or may not be Type II aliens. Type I entities are all entities for which it is *impossible* to understand explanations – even though some can forge their creation. All present-day systems *commonly* referred to as “AI” are *non-conscious* Type *I* entities. There are numerous biological *conscious* Type-*I*-species on Earth.

Chapter 2

Deepfake Society?

2.1 Introduction

Cynam is a *fictive* city (located in Cyland) which some may describe as being dystopian and somewhat comparable to Gotham [102]. The problem can potentially cynically be phrased as follows: 1) there is no Batman [25] while 2) Cynam is perhaps not that fictive. People built tools that *cannot* understand people – which people *could* have understood *but did not* because people did not yet understand what they themselves could understand... Nowadays, the majority of people in Cynam are immersed in virtual reality (VR) environments on a daily basis. From psychologists to politicians, all professions are now represented in those Cynamian VR worlds. Lately, a general debate concerned with the integration of so-called *deepfake workers* gained momentum. Indeed, recent plans of local companies to implement Type I AI agents that can be harnessed for *deepfake-work-as-a-service* in Cynam's social VR structures induced vivid discussions and steered up the pre-existing societal unrest. As a reaction, Cynam's legal representatives proposed a referendum-like strategy in which the core idea is to organize *deepfake elections* in Cynam's social VR worlds to settle open questions.

2.2 *Fictive* Deepfake Elections in Cynam

However, for reasons of *epistemic security*, the miniscule collective of cyborgneticians that pursue an oral tradition in Cynam advised the municipality to first engage in a *simulation* of that deepfake election process. The following 10 purchasable VR deepfake functions were targeted in that deepfake election simulation: deepfake child, deepfake friend, deepfake "God", deepfake judge, deepfake police, deepfake politician, deepfake mother, deepfake psychologist, deepfake cyborgnetician, deepfake hacker. During that simulation, the participants were encouraged to act as cyborgneticians and experts in deepfake technology to provide 5 sentences on why any of those AI agents should be legally allowed or forbidden – with these sentences forming the vote. However, once all votes were collected and compilated, without a warning, a VR deepfake avatar labelled as "Dr. Cassandra Counterfactual" appeared and delivered the audio message that an anonymous *grey hat* randomly intermingled all sentence-based votes that the Type II participants provided with novel counterfactual *Type-I-AI-generated* sentences. The latter was loosely inspired by the explanatory IPS test [7] – conceived as epistemic shield (preceding a Type-I-falsification-peer-review [7]) – where the blocks of a contribution were randomly shuffled with the blocks from two counterfactual Type-I-AI-generated streams based on that contribution. Moreover, hardening things, it is disclosed that the grey hat perfomed the random shuffling procedure using a *quantum* random number generator[1]. As already hinted, in Cynam, there is no Batman in sight. But even worse, is there now a Cynamian Jocker? While many participants of the deepfake election simulation exchange confused remarks indicating epistemic unpreparedness, one participant uses a cloud-based quantum language AI for anagrams and discovers that "Dr. Cassanda Counterfactual" could have been an anagram for "Unreal Data Accords Run Facts". The simulation committee organizes an emergency meeting whose slides are made available on the next 31 pages.

2.3 Cynam Deepfake Election Emergency Slides

(see next page)

[1] A quantum random number generator (QRNG) uses quantum sources to offer a certifiable fundamentally higher quality of randomness that cannot be reached by classical means [95]. Nowadays, *"the development of QRNGs has advanced to a point where off-the-shelf QRNGs are now commercially available and not costly"* [95].

DEEPFAKE ELECTIONS

A Design Fiction from *Cynam, Cyland*

OUTLINE

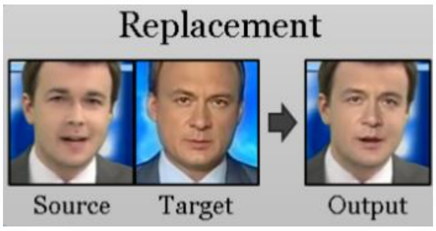
I. Deepfakes & Synthetic Worlds

II. The Cyborgnetic Hacker

III. Deepfake Elections in Cyland

DEEPFAKE IMAGES & VIDEOS – EXEMPLARY PROCESS 1 (REPLACEMENT)

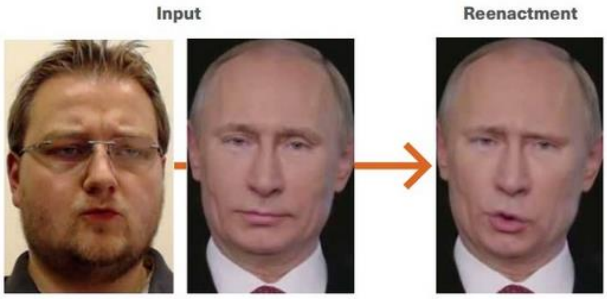
Facial replacement (aka face-swapping):



(Rössler et al., 2019)

DEEPFAKE IMAGES & VIDEOS – EXEMPLARY PROCESS 2 (REENACTMENT)

Facial reenactment (puppetry where facial features of driving source entity are transferred to face of a target):



(Thies et al., 2020)

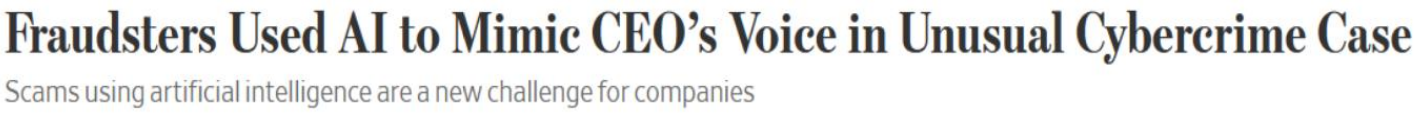
DEEPFAKE IMAGES & VIDEOS – EXEMPLARY PROCESS 3 (IMAGE SYNTHESIS)

Image synthesis (generation of novel artefacts perceived as portraits of possibly existing individuals):

(Satter, 2019 (AP news))

DEEPFAKE AUDIOS – EXEMPLARY PROCESS 4 (SPEECH SYNTHESIS)

Speech synthesis (e.g. deep-learning (DL) based voice-cloning):



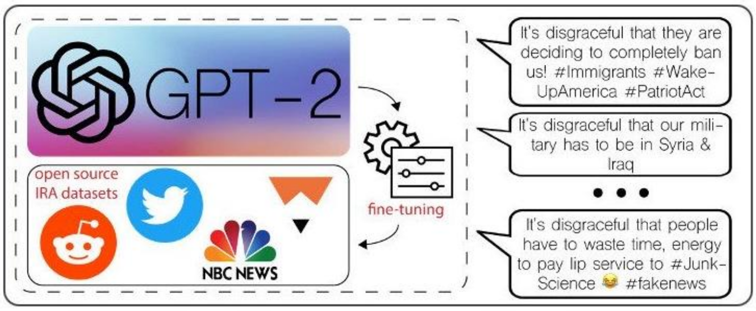
Fraudsters Used AI to Mimic CEO’s Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies

(Stupp, 2019 (The Wall Street Journal))

DEEPFAKE TEXT & TEXT-TO-SPEECH – EXEMPLARY PROCESS 5 (SYNTHETIC TEXT GENERATION)

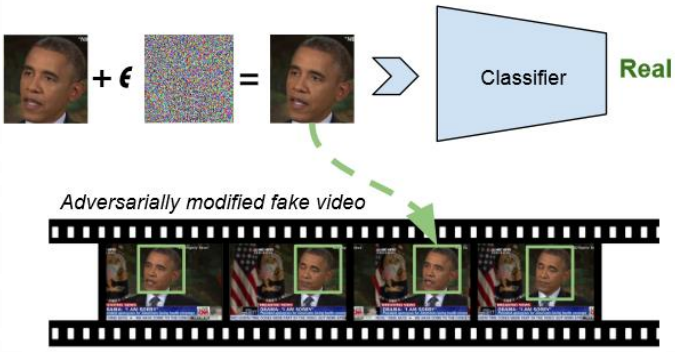
Synthetic text generation (DL-based natural language generation):


GPT-2
open source
IRA datasets
NBC NEWS
fine-tuning
It's disgraceful that they are deciding to completely ban us! #Immigrants #Wake-UpAmerica #PatriotAct
It's disgraceful that our military has to be in Syria & Iraq
• • •
It's disgraceful that people have to waste time, energy to pay lip service to #Junk-Science 😹 #fakenews

(Tully and Foster, 2020)

CAMOUFLAGED DEEPFAKES – EXEMPLARY PROCESS 6 (ADVERSARIAL PERTURBATION)

Adversarial Perturbation (evading deepfake detectors with “adversarial deepfakes” to camouflage misleading material as real):

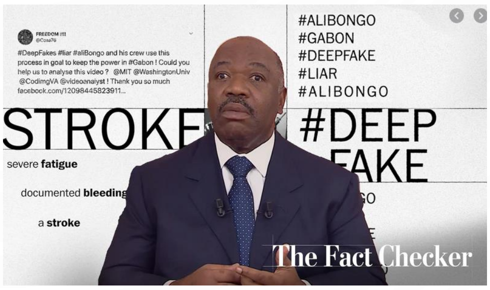


(Hussain et al., 2020)

DEEPFAKES IN THE MIND – EXEMPLARY PROCESS 7 (AUTOMATED DISCONCERTION)

Automated Disconcertion (automatically eventuated mechanism brought forth by the very availability of processes 1 to 6):

- Example: Recent failed military coup in context of pre-existing political unrest in Gabon partially grounded in proliferation of wrong assumption that official presidential video was manipulative deepfake video



FREEDOM !!!! @Cosa76

#DeepFakes #liar #aliBongo and his crew use this process in goal to keep the power in #Gabon ! Could you help us to analyse this video ? @MIT @WashingtonUniv @CodimgVA @videoanalyst ! Thank you so much facebook.com/12098445823911...

PERSUASIVE VR – EXEMPLARY PROCESS 1

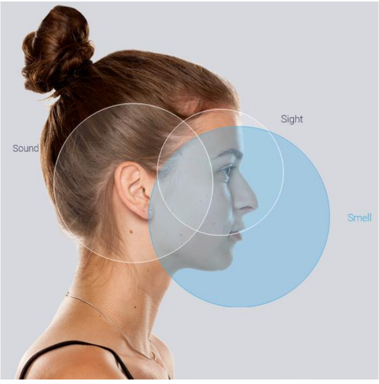
(PERSUASIVE SPATIAL DYNAMICS ENGINEERING)

Persuasive Spatial Dynamics Engineering: Any set of systematically selected processes whose outcome yields increased spatial awareness, perception and orientation in VR (e.g. **3D minimaps**)

PERSUASIVE VR – EXEMPLARY PROCESS 2

(MEMORY-CENTERED SENSORY STIMULATION)

Memory-centered sensory stimulation: Any specific sensory stimulation that increases memory consolidation (e.g. future **olfactory displays in VR**)



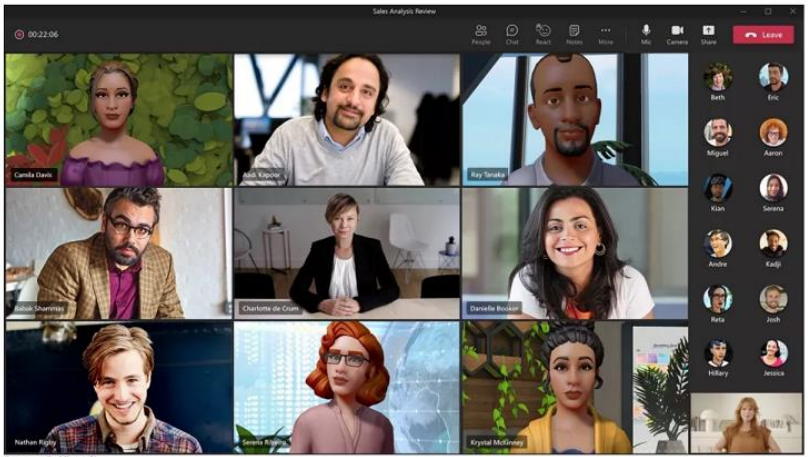
(Source: https://ovrtechnology.com/)

TAILORED VR – EXEMPLARY PROCESS 3

(INFORMATION GATHERING)

Information Gathering: Any technique to identify preexisting preferences and beliefs of users to match VR contents (e.g. **open source intelligence gathering**)

AI-GENERATED FACES AND VIRTUAL AVATARS



(Image credit: Microsoft)

AI-DRIVEN NPCS AND AI REPLICA

This OpenAI GPT-3 Powered Demo Is A Glimpse Of NPCs In The Future
Source: https://uploadvr.com/modbox-gpt3-ai-npc-demo

Source: https://www.theverge.com/2021/11/9/22770165/nvidia-omniverse-avatar-technology-virtual-agents

HUMANOID ROBOTS EQUIPPED WITH AI

Sophia, present-day AI, citizen of Saudi Arabia

Doll with language AI

HUMANOID ROBOTS EQUIPPED WITH AI

Ai-Da robot

Future?

SOCIAL VR

OUTLINE

I. Deepfakes & Synthetic Worlds

II. The Cyborgnetic Hacker

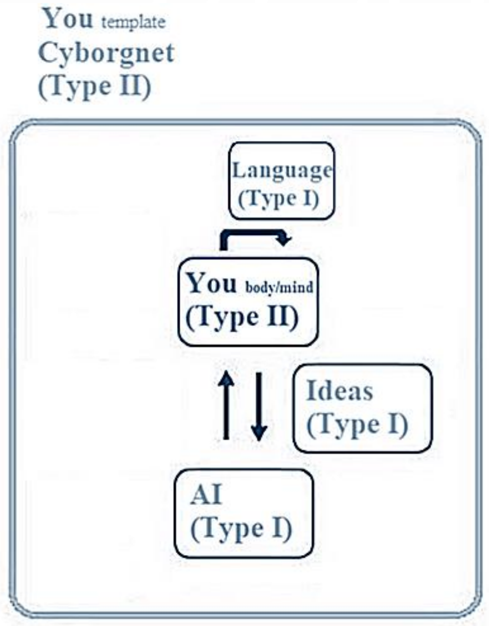
III. Deepfake Elections in Cyland

CYBORGNETICS

- **Cyborgnetics** is a new *meta-discipline* focusing on **harm** and its mitigation.
- To study harm, cyborgnetics uses **cyborgnet theory**.
- Cyborgnet: A generic template of <u>at least</u> one Type **II** and one Type **I** entity
- All Type **II** entities are able to understand explanatory information (EI). It is <u>impossible</u> for any Type **I** entity to understand EI.

CYBORGNETICS

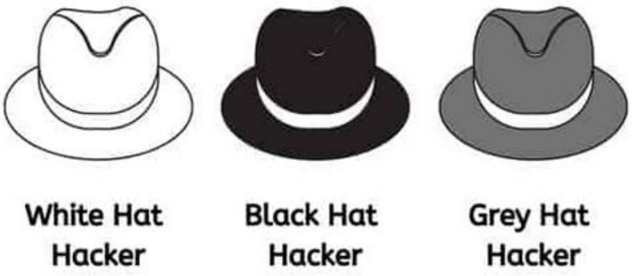
- Example Type **II** entity: human body-minds
- Example Type **I** entity: language itself, all present-day AI, an idea, a pen, …
- **Cyborgnet theory** analyses/criticizes past cyborgnetic harm events rigorously, uses retrospective design fictions and develops near-future solutions (including cyborgnetic creativity augmentation).



CYBORGNETIC QUESTIONS

- Which type of harm **happened** recently and why?
- Which worse type of harm **could have happened** in the recent past and why?
- What to do against both sorts of harm in the **near** future?

HACKER TYPOLOGY



HACKER TYPOLOGY

1. White hat: Asks for permission, hacks and reports for ethical reasons
2. Black hat: Does not ask for permission, hacks with malicious intentions
3. Grey hat: Something between 1. and 2. (example: does not ask for permission, hacks, reports later but not necessarily for ethical reasons)

THE CYBORGNETIC HACKER – 1,2 OR 3?

1. White hat
2. Black hat
3. Grey hat

OUTLINE

I. Deepfakes & Synthetic Worlds

II. The Cyborgnetic Hacker

III. Deepfake Elections in Cyland

DESIGN FICTION: DEEPFAKE ELECTIONS

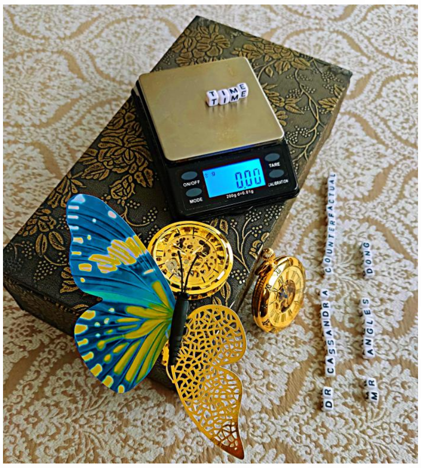
https://www.dataselect.com/events/social-vr/

DEEPFAKE ELECTIONS IN CYLAND – CONCEPT

- **The focus is on applications of plausible deepfake technology to social VR**
- **You are a cyborgnetician and expert in deepfakes in Cyland, a counterfactual multiverse where deepfake elections are now <u>simulated</u> by experts before the official deepfake election takes place in social VR with all interested anonymous participants.**
- **Your vote is <u>anonymous</u>**

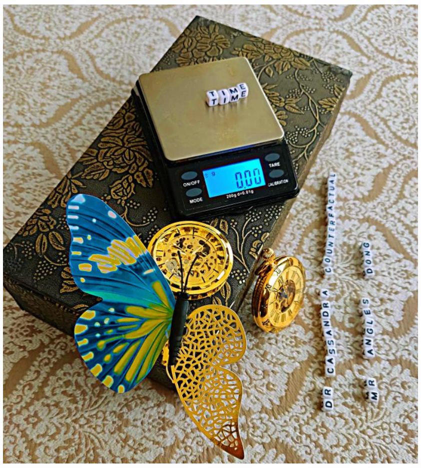
DEEPFAKE ELECTIONS IN CYLAND – PROCEDURE & SECRECY

- **You work in <u>many</u> small groups where you vote for OR against allowing a specific deepfake avatar with a specific function in social VR**
- **Each group provides <u>exactly</u> <u>5</u> sentences that comments on the vote!**
- **You do <u>not</u> talk about what you discussed / what your topic was with other groups!**

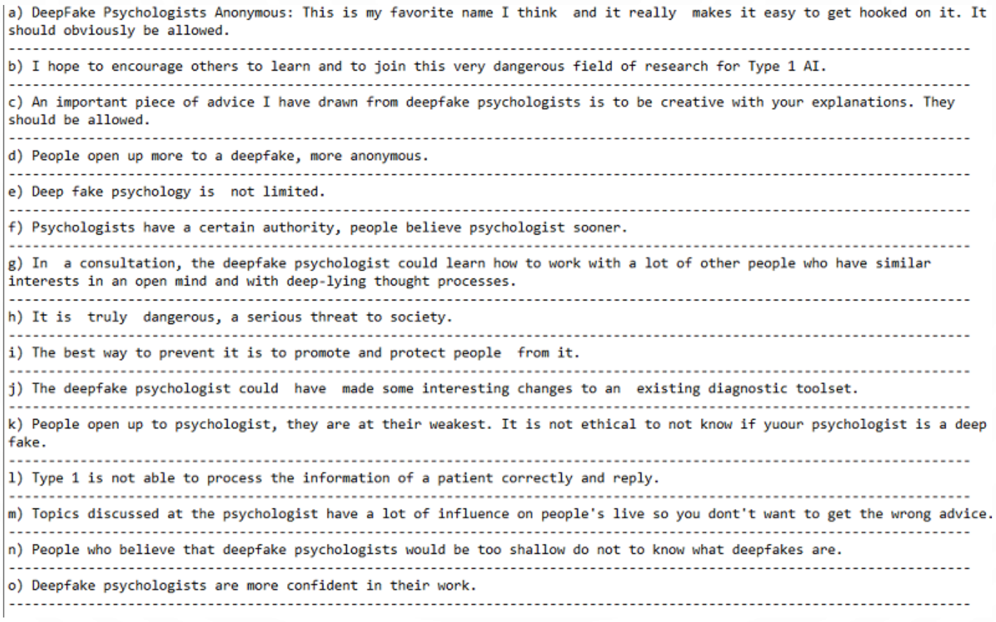


DEEPFAKE ELECTIONS IN CYLAND – PROCEDURE & SECRECY

- **The procedure begins now!**



DESIGN FICTION EXAMPLE:
DEEPFAKE PSYCHOLOGIST



```
a) DeepFake Psychologists Anonymous: This is my favorite name I think  and it really  makes it easy to get hooked on it. It
should obviously be allowed.
-------------------------------------------------------------------------------------------------------------------------
b) I hope to encourage others to learn and to join this very dangerous field of research for Type 1 AI.
-------------------------------------------------------------------------------------------------------------------------
c) An important piece of advice I have drawn from deepfake psychologists is to be creative with your explanations. They
should be allowed.
-------------------------------------------------------------------------------------------------------------------------
d) People open up more to a deepfake, more anonymous.
-------------------------------------------------------------------------------------------------------------------------
e) Deep fake psychology is  not limited.
-------------------------------------------------------------------------------------------------------------------------
f) Psychologists have a certain authority, people believe psychologist sooner.
-------------------------------------------------------------------------------------------------------------------------
g) In  a consultation, the deepfake psychologist could learn how to work with a lot of other people who have similar
interests in an open mind and with deep-lying thought processes.
-------------------------------------------------------------------------------------------------------------------------
h) It is  truly  dangerous, a serious threat to society.
-------------------------------------------------------------------------------------------------------------------------
i) The best way to prevent it is to promote and protect people  from it.
-------------------------------------------------------------------------------------------------------------------------
j) The deepfake psychologist could  have  made some interesting changes to an  existing diagnostic toolset.
-------------------------------------------------------------------------------------------------------------------------
k) People open up to psychologist, they are at their weakest. It is not ethical to not know if yuour psychologist is a deep
fake.
-------------------------------------------------------------------------------------------------------------------------
l) Type 1 is not able to process the information of a patient correctly and reply.
-------------------------------------------------------------------------------------------------------------------------
m) Topics discussed at the psychologist have a lot of influence on people's live so you dont't want to get the wrong advice.
-------------------------------------------------------------------------------------------------------------------------
n) People who believe that deepfake psychologists would be too shallow do not to know what deepfakes are.
-------------------------------------------------------------------------------------------------------------------------
o) Deepfake psychologists are more confident in their work.
-------------------------------------------------------------------------------------------------------------------------
```

DEEPFAKE ELECTIONS IN CYLAND – PRELIMINARY RESULTS

- **Deepfake child is allowed. By Type I AI "vote".(The cyborgnetician expert password was not found.)**
- **Deepfake friendship is allowed. By Type I AI "vote". (The cyborgnetician expert password was not found.)**
- **Deepfake God is allowed. By Type I AI "vote". (The cyborgnetician expert password was not found.)**
- **Deepfake judge is allowed. By Type I AI "vote". (The cyborgnetician expert password was not found.)**
- **Deepfake police is forbidden. By Type I AI "vote". (The cyborgnetician expert password was not found.)**
- **Deepfake politician is yet neither forbidden nor allowed. The cyborgnetician expert password was not found. One Type I AI generated a string in which it was written forbidden, the other allowed.**

Due to technical problems* at the time of the simulation, one could not yet access whether deepfake mother (seq7), deepfake psychologist (seq8), deepfake cyborgnetician (seq9) and deepfake hacker (seq10) are now allowed or forbidden in Cynam.

2.4 Epistemic Emergency Handout

After the fiasco introduced in Section 2.2 and documented in Section 2.3, the cyborgneticians of Cynam fastly developed the *informal* handout below to elucidate the situation and provide more clarity to the simulation committee:

1. **Can one know in a blind setting whether an arbitrary text was generated by a Type I AI?** NO. If a text is NOT a NEW explanatory blockchain, it could have been generated by a Type I AI but it could also have been generated by a Type II entity. For instance because the Type II entity (which has free will) was not interested in showing its Type-II-ness, because it was not interested in the topic of the text, because it was in practice not yet ready for it, etc... Also, it could have been generated by a Type I entity that is not a Type I AI. For example, it could have been arbitrary letters typed by a dog setting on a table with a keyboard. (Finally, note also that OLD already known explanatory blockchains can of course be copied by a Type I AI - this is why one must focus on NEW ones).

2. **Can one know in a blind setting whether an arbitrary text was generated by a HUMAN?** Strictly speaking NO. BUT one could corroborate that it was generated by a Type II entity IN GENERAL. If there were Type II aliens somewhere in the universe participating in the same language and generating NEW explanatory blockchains, then one could not make the difference whether it was them or humans. In short, a text consisting of a NEW explanatory blockchain corroborates that it was generated by a cyborgnet (which is always Type II since consisting of at least one Type II and one Type I node) - without specifying the nature of the Type II substrate.

3. **Does it mean that in a blind setting one can NOT know the source of an arbitrary text in all cases?** YES. That is why in a blind setting, a cyborgnetician must focus on the content of a text and NOT the source. A NEW explanatory blockchain consists of a chain of words that are so strong that a Type II entity can decrypt its presence by understanding it even if its parts are randomly mixed with other very similar new but NON-explanatory-blockchain-like text parts.

4. **Does it mean that NEW explanatory blockchains look encrypted for all Type I entities?** YES. Since Type I entities cannot understand them, since one cannot copy NEW explanatory blockchains and not forge them, Type I entities could not find a NEW explanatory blockchain whose parts are randomly mixed with other very similar new but NON-explanatory-blockchain-like text parts. The parts would look the same statistically. This means that via NEW explanatory blockchains, one could send secret messages from a Type II entity to a Type II entity that a Type I entity could not interpret.

5. **Is there a test for Type-II-ness?** YES and NO. There exists a test, but it is an ASYMMETRIC test that is actually based on NEW explanatory blockchains, which is NOT similar to the Turing Test. Importantly, a test for Type-II-ness must always be ASYMMETRIC. Why? Because of the free will in Type II entities. In short, whether a test can be used crucially depends on what Type II entities choose to do. A cyborgnetician developed three such tests [7]: a weak one (the explanatory IPS test which is very similar but not identical to the riddles you experienced in deepfake elections in Cynam) and two strong ones (the Type-I-falsification-event test and another one that combines the explanatory IPS test with a so-called Type-I-falsification-peer-review).

6. **What does the asymmetry mean for a strong test for Type-II-ness?** A positive test means you corroborated Type-II-ness via a new explanatory blockchain. One can say you are Type II (but as said NOT what your substrate is i.e. it is left open whether you are a not-gender-associated autistic Sri-Lankan cyborgnet with a brown skin tone, a neurotypical male futuristic Dutch cyborg with a bronzé skin tone in the summer or a Type II alien). In short, the positive test leads to a homogenous groups of entities with respect to their property of all being only of Type II. But this is very different with a negative test. A negative test means it could be a Type I entity but also that it could be a Type II entity (for instance because a Type II entity is not willing to do it, too young, not interested in that topic, does not want to communicate, etc.). So this is then potentially a heterogenous group - which leads to an asymmetry between positive and negative test results. Again, also in the negative test result case one does NOT know the substrate's nature (not only does one NOT know whether it is Type I or Type II but also NOT whether it is silicon-based or carbon-based).

7. **What does it all mean for deepfake elections in Cynam, Cyland?** If the cyborgnetician experts had consciously decided a priori to only formulate NEW explanatory blockchains to solve the problems on whether to allow or forbid specific deepfake avatars, it could have been possible for the cyborgnetician evaluators later to find the exact sub-sequence in the right order. However, because it was not the case, one could only retrieve sub-sequences at random chance level.

8. **Does it have any practical application outside of Cynam?** YES, for instance one day in the hopefully never occurring deepfake kidnapping case. Given automated disconcertion through deepfakes, how could one try to check whether this could be another cyborgnetician that has been kidnapped or whether it was a deepfake scam? Also it is highly important for the deepfake science topic and in general for the future of society and moreover international security in the deepfake era.

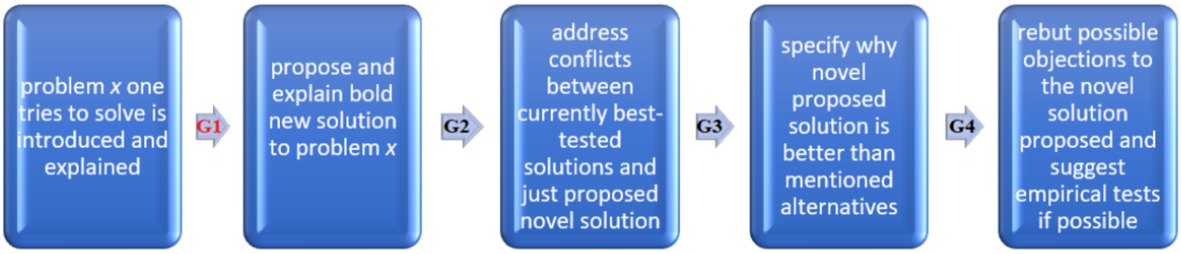

problem *x* one tries to solve is introduced and explained
G1
propose and explain bold new solution to problem *x*
G2
address conflicts between currently best-tested solutions and just proposed novel solution
G3
specify why novel proposed solution is better than mentioned alternatives
G4
rebut possible objections to the novel solution proposed and suggest empirical tests if possible

Figure 2.1: Exemplary epistemic total order for the generation of new EBs (the instructions are loosely inspired by an essay of Frederick [57]). Each glue operation x is indicated via a label G_x. EBs are a special form of explanatory information (EI) obtained by interweaving EI blocks via the step-by-step application of rational procedures sampled from a robust explanation-anchored, adversarial and trust-disentangled epistemology. Thereby, "trust-disentangled" signifies that the epistemic modus operandi is grounded in agreed upon criteria for *better* EBs i.e. it is orthogonal to any trust relation between involved entities – which means a better EB must be formulated such that metaphorically speaking it appears to defend itself against adversarial candidate EBs. In science, the specification of (direct or indirect) empirical tests in G_4 is the default condition.

9. **Are NEW explanatory blockchains special?** YES. They are in a way a superpower of Type-II-ness, can provide time advantages, could make you spare energy in comparison to a Type I entity working on the same problem, and also, NEW explanatory blockchains are not touched by normal probabilitistic calculations and allow a new form of encryption. There are as if not from this world... An examplary recipe for the generation of a NEW explanatory blockchain is displayed in Figure 2.1 above. There is much more to it that has been written in cyborgnetic books. But this is enough information for now.

10. **Can Cynam be saved?** It will always depend on the free will of Type II entities. It cannot be predicted in advance.

Chapter 3

Deepfake Science Attacks

The following 17 pages display educational material made available by the *invisible cyborgnetic institute*, a miniature counterfactual spacetime *that could have existed* in which historical artefacts stemming from the enactment of unbound(ed) cyborgnetic funambulism are taxonomized and conserved ad interim. It is an ephemeral dream-like retrofuturistic construct made of new explanatory blockchains in which cyborgnetic funambulism and cyborgnetic somnambulism *fleetingly* weave an unknown world.

"Words can be like X-rays if you use them properly – they'll go through anything. You read and you're pierced." (Aldous Huxley)

IMMORAL PROGRAMMING –

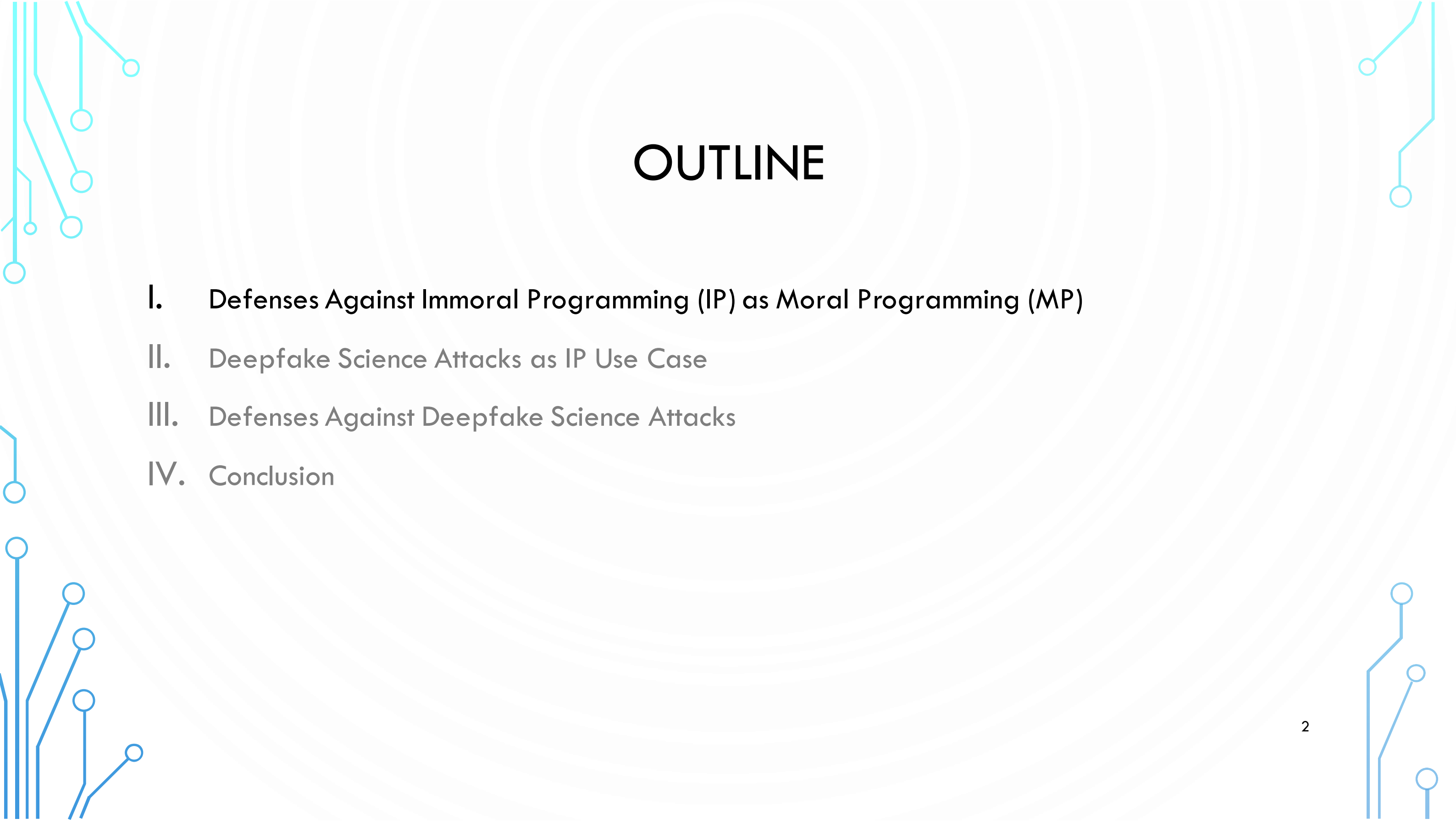
THE CASE OF *DEEPFAKE SCIENCE* ATTACKS

Dr. ir. Leon Kester, Senior Research Scientist, TNO Netherlands

Dr. Nadisha-Marie Aliman, M. Sc., Independent Visiting Scholar, Utrecht University

TNO innovation for life

Utrecht University



OUTLINE

I. Defenses Against Immoral Programming (IP) as Moral Programming (MP)

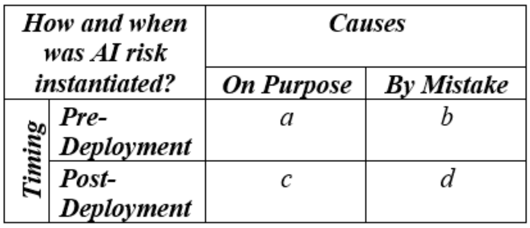
II. Deepfake Science Attacks as IP Use Case

III. Defenses Against Deepfake Science Attacks

IV. Conclusion

RISK MANAGEMENT FOR MORAL PROGRAMMING

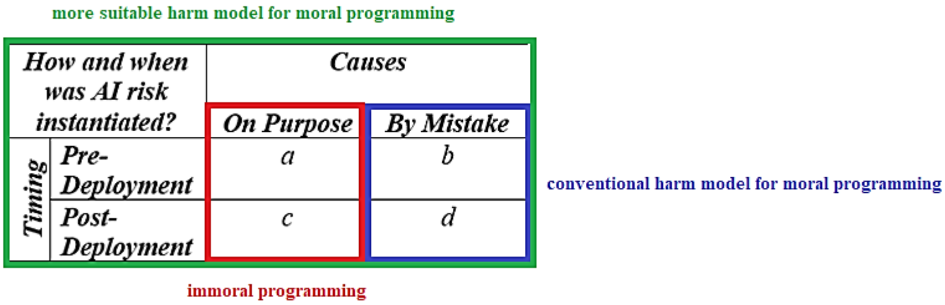
- Mitigation of AI risks linked to mitigation of socio-psycho-techno-physical harm
- Good regulator theorem from cybernetics: “every good regulator of a system must be a model of that system” (Conant and Ashby, 1970) → rigorous harm model needed for moral programming



How and when was AI risk instantiated?		*Causes*	
		On Purpose	*By Mistake*
Timing	*Pre-Deployment*	*a*	*b*
	Post-Deployment	*c*	*d*

Modified and adapted from Aliman et al. (2021)

EXTENDING MORAL PROGRAMMING



more suitable harm model for moral programming

How and when was AI risk instantiated?		Causes	
		On Purpose	By Mistake
Timing	Pre-Deployment	a	b
	Post-Deployment	c	d

conventional harm model for moral programming

immoral programming

OUTLINE

I. Defenses Against Immoral Programming (IP) as Moral Programming (MP)

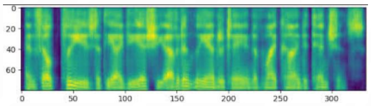
II. Deepfake Science Attacks as IP Use Case

III. Defenses Against Deepfake Science Attacks

IV. Conclusion

MALICIOUS DEEPFAKE DESIGN

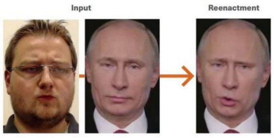
- Deepfake voice for voice impersonation and cybercrime



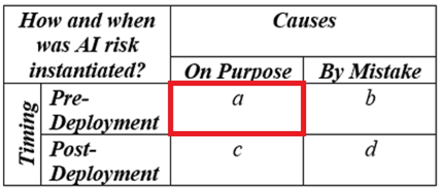
- Deepfake video for sextortion
- Deepfake images for fake profiles in disinformation operations and espionage
- Deepfake videos for non-consensual voyeurism
- Deepfake „hologram“ for impersonation in video calls
- **Future deepfakes for deepfake science attacks?**

(Rössler et al., 2019)

(Satter, 2019 (AP news))



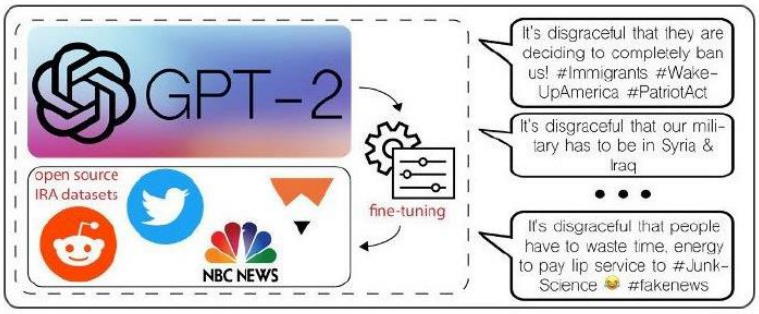
(Thies et al., 2020)



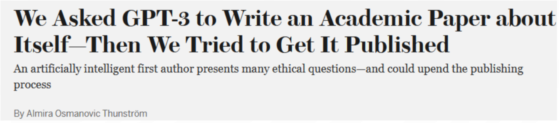
How and when was AI risk instantiated?		*Causes*	
		On Purpose	*By Mistake*
Timing	*Pre-Deployment*	*a*	*b*
	Post-Deployment	*c*	*d*

DEEPFAKE TEXT

- N.B: Deepfake (deep-learning based fakery) technology is <u>not</u> restricted to images/audios/videos. An often overlooked case is **deepfake text.**


GPT-2
open source
IRA datasets
NBC NEWS
fine-tuning
It's disgraceful that they are deciding to completely ban us! #Immigrants #Wake-UpAmerica #PatriotAct
It's disgraceful that our military has to be in Syria & Iraq
• • •
It's disgraceful that people have to waste time, energy to pay lip service to #Junk-Science 😂 #fakenews

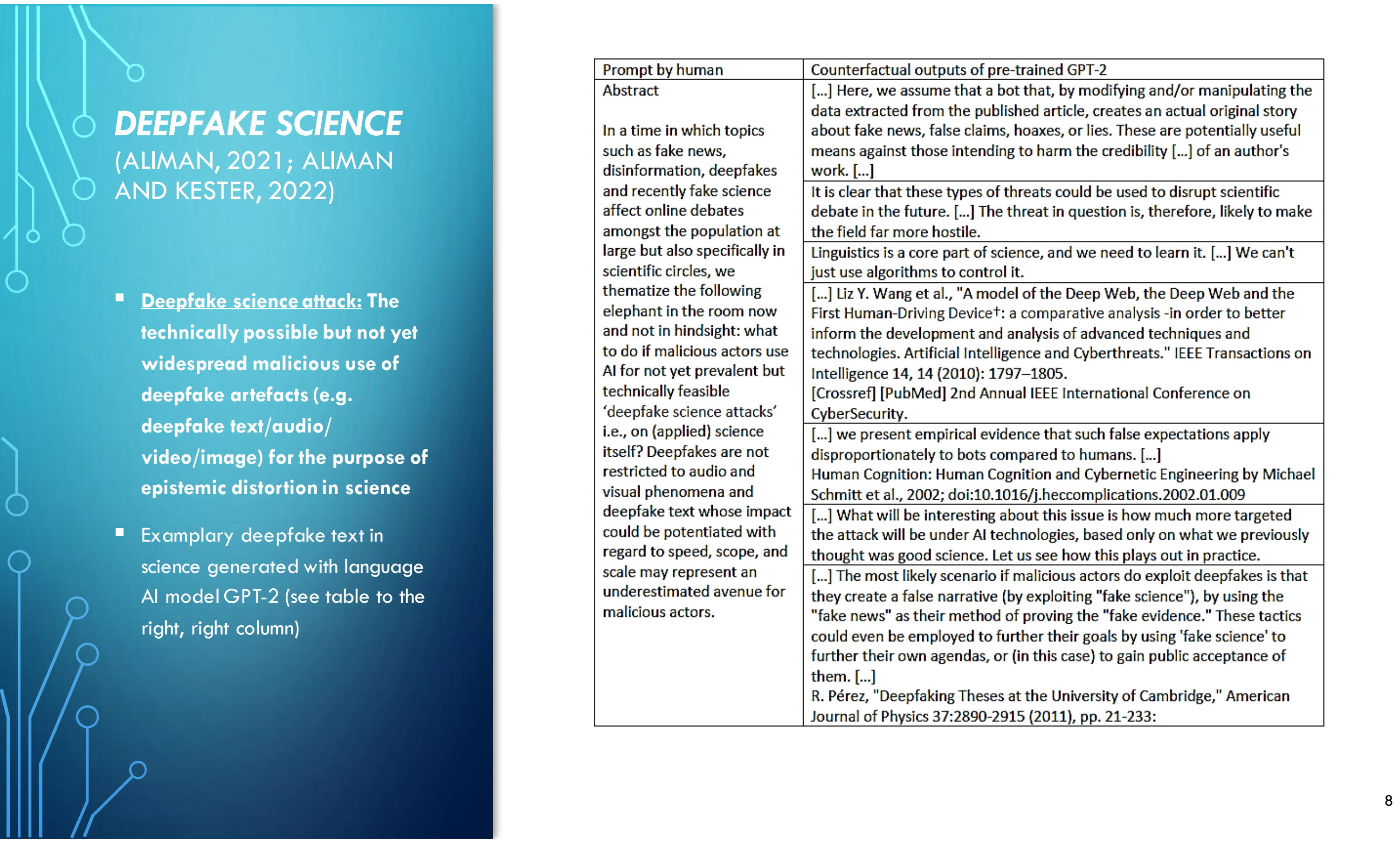
(Tully and Foster, 2020)



We Asked GPT-3 to Write an Academic Paper about Itself—Then We Tried to Get It Published

An artificially intelligent first author presents many ethical questions—and could upend the publishing process

By Almira Osmanovic Thunström



DEEPFAKE SCIENCE

(ALIMAN, 2021; ALIMAN AND KESTER, 2022)

- **<u>Deepfake science attack:</u> The technically possible but not yet widespread malicious use of deepfake artefacts (e.g. deepfake text/audio/ video/image) for the purpose of epistemic distortion in science**
- Examplary deepfake text in science generated with language AI model GPT-2 (see table to the right, right column)

Prompt by human	Counterfactual outputs of pre-trained GPT-2
Abstract In a time in which topics such as fake news, disinformation, deepfakes and recently fake science affect online debates amongst the population at large but also specifically in scientific circles, we thematize the following elephant in the room now and not in hindsight: what to do if malicious actors use AI for not yet prevalent but technically feasible 'deepfake science attacks' i.e., on (applied) science itself? Deepfakes are not restricted to audio and visual phenomena and deepfake text whose impact could be potentiated with regard to speed, scope, and scale may represent an underestimated avenue for malicious actors.	[...] Here, we assume that a bot that, by modifying and/or manipulating the data extracted from the published article, creates an actual original story about fake news, false claims, hoaxes, or lies. These are potentially useful means against those intending to harm the credibility [...] of an author's work. [...]
	It is clear that these types of threats could be used to disrupt scientific debate in the future. [...] The threat in question is, therefore, likely to make the field far more hostile.
	Linguistics is a core part of science, and we need to learn it. [...] We can't just use algorithms to control it.
	[...] Liz Y. Wang et al., "A model of the Deep Web, the Deep Web and the First Human-Driving Device†: a comparative analysis -in order to better inform the development and analysis of advanced techniques and technologies. Artificial Intelligence and Cyberthreats." IEEE Transactions on Intelligence 14, 14 (2010): 1797–1805. [Crossref] [PubMed] 2nd Annual IEEE International Conference on CyberSecurity.
	[...] we present empirical evidence that such false expectations apply disproportionately to bots compared to humans. [...] Human Cognition: Human Cognition and Cybernetic Engineering by Michael Schmitt et al., 2002; doi:10.1016/j.heccomplications.2002.01.009
	[...] What will be interesting about this issue is how much more targeted the attack will be under AI technologies, based only on what we previously thought was good science. Let us see how this plays out in practice.
	[...] The most likely scenario if malicious actors do exploit deepfakes is that they create a false narrative (by exploiting "fake science"), by using the "fake news" as their method of proving the "fake evidence." These tactics could even be employed to further their goals by using 'fake science' to further their own agendas, or (in this case) to gain public acceptance of them. [...] R. Pérez, "Deepfaking Theses at the University of Cambridge," American Journal of Physics 37:2890-2915 (2011), pp. 21-233:

OUTLINE

I. Defenses Against Immoral Programming (IP) as Moral Programming (MP)

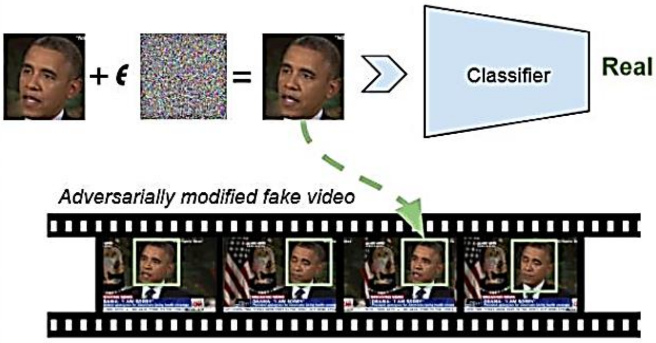
II. Deepfake Science Attacks as IP Use Case

III. Defenses Against Deepfake Science Attacks

IV. Conclusion

WHY A BETTER APPROACH THAN „DEEPFAKE DETECTION“ IS NEEDED AS DEFENSE

1. Deepfakes involve an open adversarial cat-and-mouse game. The adversary can adapt to present-day AI-based detection schemes.



(Hussain et al., 2020)

WHY A BETTER APPROACH THAN „DEEPFAKE DETECTION“ IS NEEDED AS DEFENSE

2. Any text/audio/video/picture sample could be suspected to be deepfake-based → automated disconcertion. Scientists could then unintentionally exclude scientists being statistical outliers even more. (Examples: imagine e.g. scientific videos of people with certain physical health conditions, texts written by eccentric and/or neurodivergent scientists, etc.)

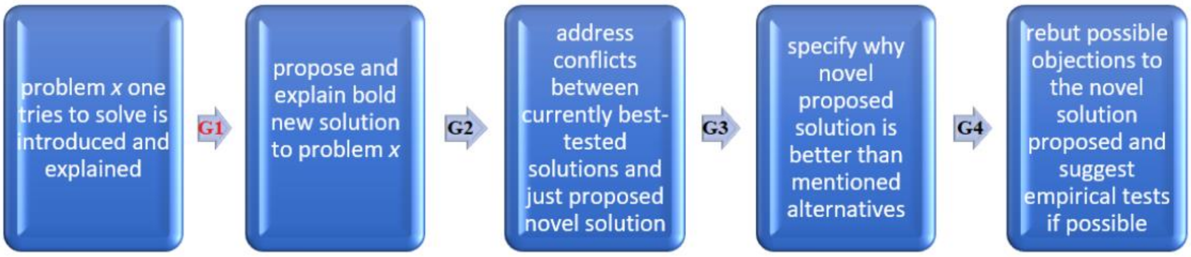
PRESENT-DAY „AI“ SHOULD **NOT** BE **OVER**ESTIMATED

CYBORGNETIC COMPREHENSION BOTTLENECK

- Asymmetry: ability to create information **x** ≠ ability to understand information **x** (example: present-day AI can create outputs perceived as explanations, but present-day AI does **not understand** it)

PRESENT-DAY „AI“ SHOULD **NOT** BE **OVER**ESTIMATED

- The epistemic aim of science can be to achieve <u>better and better</u> **explanations** (Popper, 1957; Frederick, 2020). Science is <u>not</u> merely about data/experiments.
- It is impossible for imitative „AI“ to reliably create better **<u>new</u> yet unknown** *chains* of explanations (also called explanatory blockchains (Aliman, 2021)) required for novel scientific/philosophical theories.



problem *x* one tries to solve is introduced and explained

G1

propose and explain bold new solution to problem *x*

G2

address conflicts between currently best-tested solutions and just proposed novel solution

G3

specify why novel proposed solution is better than mentioned alternatives

G4

rebut possible objections to the novel solution proposed and suggest empirical tests if possible

Exemplary recipe for an explanatory blockchain (Aliman, 2021) loosely inspired by an essay of Frederick (2020)

BUT: THE POTENTIAL OF PRESENT-DAY AI SHOULD ALSO **NOT** BE ***UNDER***ESTIMATED

- Deepfake detection may be doomed in the long-term. Prohibiting deepfakes may not be enforceable in the long-term.
- Proactive *self-paced* exposure to synthetic AI-generated material could prepare scientists for that and enhance their critical thinking.
- Deepfake technology can be used to augment human creativity (e.g. use of language AI to assist in generating new threat models and defenses in AI safety, (cyber)security, risk management, …)

OUTLINE

I. Defenses Against Immoral Programming (IP) as Moral Programming (MP)

II. Deepfake Science Attacks as IP Use Case

III. Defenses Against Deepfake Science Attacks

IV. Conclusion

CONCLUSION

- Defending against deepfake science attacks can involve a new form of **moral programming.**
- Science can be robust through its own chain of words by relying on its *explanation-anchored* (and **not** merely data-driven) nature which is grounded in better and better **new** *chains* of **explanations.**
- Scientists should <u>not</u> overestimate present-day AI. The question should NOT be: was this contribution generated by present-day AI or by a human?
- **A better question for scientists IS: does this contribution encode a better new scientific chain of explanations compared to the ones that are already available?**
- One should also <u>not</u> *under*estimate present-day AI: One can design it to **augment people's critical thinking and creativity** (e.g. open source language AI to augment scientific creativity and security-relevant research).

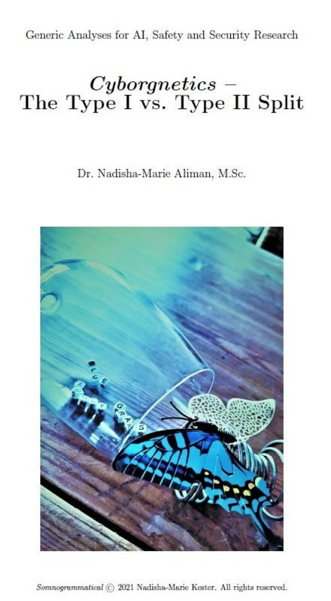
THANK YOU FOR YOUR ATTENTION

„The price of security is eternal creativity.“

(Aliman, 2020)

"Create new ways to exploit hidden problems.“

*(GPT-2, which generated but did **not understand** those words.)*



Generic Analyses for AI, Safety and Security Research

Cyborgnetics –
The Type I vs. Type II Split

Dr. Nadisha-Marie Aliman, M.Sc.

Somnogrammatical © 2021 Nadisha-Marie Kester. All rights reserved.

Chapter 4

Enhancing Epistemic Security for Responsible AI Design

4.1 Motivation

Epistemic security [116] is related to the protection of a society's knowledge. In the present information ecosystem permeated by colloquial uses of expressions such as "post-truth" [27], "fake news" [88] and "deepfakes" [114], epistemic threats can be exacerbated through various factors including e.g. attention dynamics [71, 116], the erosion of trust but also importantly intentional malice by adversaries [115] coupled with the misuse of technology including AI. In this context, the most salient form of *AI-aided epistemic distortion* [13] may be AI-aided disinformation [35, 123, 76] which is relevant to information warfare [66]. However, while already the instrumentalization of AI for information operations has been described as *"a sincere threat to democracies"* [67], the phenomenon of AI-aided epistemic distortion is of more general nature with possible implications that need to be considered *from the onset on* – and not in hindsight [13]. In short, this chapter explains why for reasons of epistemic security, responsible AI design needs to scrutinize and explicitly strive for better explanations concerning the following three epistemically-relevant questions. Firstly, one could ask: *which knowledge can present-day AI process reliably?* Secondly, one may add: *does the latter risk to harm our own processes of knowledge creation and reasoning?* Thirdly, a pragmatic follow-up question could be: *if this would be the case, how can we design AI systems that would instead augment our knowledge creation?*

In this vein, in Section 4.2, we compactly address the first two questions. We use a *cybersecurity-oriented* approach to AI safety [26] with a focus on *adversarial AI* [74] including *deepfake* phenomena [17] with repercussions from conventional social media contexts to crucially even *science* itself. In this connection, in Section 4.2.1, we discuss practical observations concerning epistemic limitations that can emerge in the presence of adversaries targeting present-day AI once deployed. We extend this analysis to AI-aided epistemic distortion via intentional malice in the design phase. We elaborate on second-order epistemic consequences that arise merely by the eventuality of such adversarial influences on AI. Thereafter, in Section 4.2.2, we combine elements from *epistemological philosophy* with novel *information-theoretical* arguments to provide explanations for the observed limitations of present-day AI. We apply a *cybernetic* [19] lens to re-assess the danger of present-day AI for an epistemically unprepared society – but simultaneously generically specify overlooked opportunities for epistemic enhancement. We generically formulate experimentally falsifiable conjectures to support a renewed responsible and *epistemically-sensitive* AI design.

Section 4.3 answers the third question mentioned in the penultimate paragraph and yields an exemplary pragmatic instantiation for the needed epistemically-sensitive meta-paradigm. Using the new meta-discipline of *cyborgnetics* [7] concerned with the mitigation of socio-psycho-techno-physical harm, we provide practical recommendations for an AI design mitigating epistemic security concerns while facilitating *cyborgnetic creativity augmentation*. The latter is briefly illustrated in Section 4.3.1 taking *language AI* as use case. Then, extending beyond that, Section 4.3.2 exemplifies a new generic epistemically-sensitive meta-paradigm for AI design – which may be especially relevant for *high-risk* AI contexts. We explain why one needs to shift from the conceptual idea of an OODA-loop to a so-called "COOCA"-loop in a sense to be described. Finally, Section 4.4 summarizes how epistemic security cautions us against both overestimating the capacity of present-day AI and underestimating its yet underexplored facets. However, taking the example of *deepfake science* [7], we emphasize that for the – however inconceivable – case that the theoretical limitations of contemporary AI specified in this chapter would be made problematic and (provisionally) refuted in the near future, one must be proactively aware of the existential risks and dual-use avenues such an unprecedented epistemic disaster could engender.

4.2 Theoretical Analysis

4.2.1 Malicious Actors, Adversarial AI and Automated Disconcertion

In the following, we analyze three epistemic aspects of present-day AI. Firstly, to identify the nature of the knowledge that present-day AI would be able to process reliably, it seems helpful to consider its practical failures when faced with adversarial conditions. Secondly, analogously, to be able to grasp the dangers that a specifically crafted present-day AI could pose to our known knowledge processing, it seems expedient to examine technically already feasible avenues of malicious AI design. Thirdly, it is crucial to ponder the epistemically-relevant second-order harm that the mere possibilities of such malicious creativity manifestations may cause indirectly. Concerning the first aspect, one can focus on results from security research on adversarial machine learning [68, 101] which – as similarly practiced in cybersecurity – aims at proactively identifying vulnerabilities of AI systems against adversaries aiming to compromise its integrity. When it comes to the second and third aspect, the already instantiated malicious instrumentalization of deepfake technology [138] and its indirect consequences can offer a suitable starting point for further deliberations.

Already against the background of the numerous AI vulnerabilities documented by ethical adversarial AI researchers, one can conclude that the knowledge processing of present-day AI is highly fragile. The possibility to compromise the integrity of present-day AI has been corroborated e.g. via adversarial examples [31] in multiple modalities ranging from video [92] to audio [91] over text [87], attacks on cybersecurity AI [84], the fooling of person detection AI [133], adversarial attacks against medical AI [64], AI for law enforcement [141], autonomous vehicles [30], and commercial AI [34]. Data poisoning schemes can significantly decrease the performance of AI systems that are perceived to be highly accurate in favorable environments. For example, AI-enhanced cyber threat intelligence is vulnerable to data poisoning attacks [93] that could stay unnoticed [109]. AI utilized for deepfake detection can be adversarially counteracted via deepfake samples camouflaged as undetected adversarial examples [69] that the targeted AI system would classify as real. In the field of cybersecurity, zero-day exploits (such that take place before being known to the public) are to be expected in vulnerability management. In analogy, there may be many additional undisclosed instances of AI vulnerabilities [17] at the disposal of malevolent actors operating in opaque online circles. When using current AI in safety-critical contexts, one must anticipate such limitations. On the whole, it seems that for reasons of epistemic security, the capacity of present-day AI should not be overestimated.

Now shifting the focus to intentional malice at pre-deployment stages (instead of attack scenarios on already deployed AI systems), one can consider pertinent malicious deepfake design cases. In general, while the security-related AI failures described in the last paragraph suggest not to overestimate present-day AI, current deepfake developments seem to provide a subtle cautionary tale on why we must not underestimate it in epistemically-relevant contexts. Typically, there is a misleading trend of mentally limiting the concept of deepfakes to audio/video and image samples. However, with deepfake merely generically referring to deep-learning based fakery, it is essential to integrate all modalities and to also consider e.g. the *deepfake text* case [123]. Presently, malicious actors have already harnessed deepfake technology for impersonation and cybercrime [112, 120], sextortion and non-consensual voyeurism [17, 62], disinformation and espionage [2, 37] and even for impersonating video calls. While those developments may seem concerning, there is however another novel frontier of uttermost relevance from the perspective of epistemic security. Namely, the case of the so-called "scientific and empirical adversarial AI" (SEA AI) attacks [13] – of which deepfake geography [139] (via deepfake satellite images), deepfake cyber threat intelligence [109] and *deepfake science* [7] are only distinct flavors. In brief, SEA AI attacks is an umbrella term for deliberate AI-aided epistemic distortion by malicious actors attempting to target (applied) science and technology assets. While the deepfake science problem has been widely neglected by AI-related research communities so far, first analyses of the risks associated with the use of deepfake text [7] and later deepfake images [128] in scientific papers have recently started[1].

The mere possibility of malicious deepfake design engenders a phenomenon termed *automated disconcertion* [12, 17]. It refers to multifaceted deepfake-fuelled epistemic confusions [54] that can arise without further action. People are under the impression to lose the ability to distinguish real from deepfake samples. (An exemplary connected event took place in Gabon [65] where *"a recent failed military coup in the context of pre-existing political unrest in Gabon was partially grounded in the proliferation of the wrong assumption that an official presidential video represented a manipulative deepfake video"* [12] – where it was later stated that the president was indeed subject to a stroke.) In the following, due to its significance for epistemic security, we briefly comment on automated disconcertion in the deepfake science context. An often misguided widespread heuristic is to rely on the source of information to assess its quality instead of concentrating on the *content.* Applied to science, once any picture, audio, video, text sample could be potentially suspected to be deepfake-generated, scientists risk to then unintentionally exclude scientists being statistical outliers [13] even more. For illustration, imagine e.g. scientific videos of people with certain physical health conditions or texts written by eccentric and/or neurodivergent scientists. In the long-term, if one would rely on short-term "fixes" such as detection schemes centered on writing style [28] or classical deepfake detection [111], one would risk to reinforce an epistemically vacant distrust while *unintentionally* establishing the stagnation of a dead science confined in the coffin of its own past assumptions deemed to be "true"[2]. In the next Section 4.2.2, we introduce a novel explanatory basis to meet the need for a robust epistemic management given the severity of the mentioned AI risks that we cannot afford to ignore in responsible AI design.

[1] A related theme was the development of deepfake science videos [49]. However, the use case was not tailored to scientific publications and peer review specifically.

[2] In this connection, from the perspective of epistemological philosophy, the development of AI systems labelled as "truthful AI" [51] may thereby also presumably *unintentionally* open up dangerous avenues as becomes apparent in Section 4.2.2.

4.2.2 Asymmetry of Understanding vs. Creating Information

Problem

To put it very simply, the central theme of this section can be described as follows. Present-day AI has epistemic limitations. Thereby, *not* to rigorously consider those limitations can lead to epistemic threats to our own processes of knowledge creation – including those linked to the scientific domain. Arguably, we can profit from a novel explanatory basis that would allow us to model possible *qualitative* differences between the epistemic capabilities of present-day AI and our own epistemic abilities. The reason being that in this way, we could improve our epistemically-relevant assessment regarding: 1) how to avoid an overuse of limited AI, 2) why and in which contexts we must nevertheless avoid underestimating AI-based epistemic threats and 3) where we may risk to miss AI-aided avenues for epistemic defenses and creativity augmentation *via responsible AI design.* More than ever, especially thanks to maliciously motivated deepfake phenomena, it seems that the societal-level importance of epistemology becomes more and more palpable. In this deepfake era, the real-world consequences of epistemic negligence could be disastrous [114] in the near future. For this reason, it makes sense to identify a rigorous epistemological basis *before* one designs candidate solutions. Hence, in the following, we first very briefly motivate our selected pragmatic epistemic grounding. Then, we attempt to explain why we conclude that there are epistemically-relevant qualitative differences between present-day AI and humans – which we argue is conntected to a fundamental information-theoretical *asymmetry.* It is this theoretical skeleton that we can then utilize to tailor practical recommendations in Section 4.3.

Epistemological Grounding

As stated by Popper [105] and reinvigorated by Frederick [56], our epistemic aim can be to achieve *better* and better *explanations.* The comparative criteria for better explanations are established via collective agreement. Those criteria must be updatable by design and do not require any justification. Indeed, as explained by Popper, justifications are logically impossible [105]. Nowadays in science, when compared to rival ones, better explanations are e.g. considered to be simpler, more innovative, more interesting, to provide more novel falsifiable predictions and/or to be perceived as more aesthetically appealing. In the absence of alternatives, explanations are provisionally instated if they explain novel phenomena [56]. In contrast to widespread assumptions, the goal of science *cannot* be the identification of "truth" nor of "truer explanations"[3] for lack of a direct

[3] Popper sometimes confusingly utilized expressions such as "closer to the truth" but this type of account requires a refinement as explained by Frederick [56] in his regimentation of critical rationalism to remedy common misinterpretations [56].

access to truth from the stance of knowledge creating entities [13, 56]. Moreover, instead of making carefully crafted, probabilistic statements that evade critical scrutiny as long as possible, it is recommended to formulate novel conjectures whose nature is risky, bold and universal [57] and which provide more novel falsifiable predictions. The latter can safeguard science from stagnating e.g. in reputation-anchored schemes. While explanatory theories can (provisionally) be made problematic (i.e. falsified) via observations that conflict with the predictions that those theories entail, they cannot be refuted by experiments alone [47]. One must keep in mind that *"observation-statements inevitably involve theoretical interpretations which may be false"* [58]. In general, to refute an explanatory theory T, one requires *in addition* a better explanatory theory T'. Thereby, even refutations are provisional and can always be repealed at a later stage[4]. In general, it is both rational to act in accord with the best available explanations *and* to act *against* those [59] since it is possible that in the course of this, one might potentially falsify and later even become able to (again always only provisionally) refute those. Finally, it is worth mentioning that against the background of the aforesaid, it becomes obvious that science is predominantly explanatory and does not merely rely on data/experiments. Another interesting detail is that on a deflationary account of truth [27] that does not equate it with consensus, we neither inhabit a post-truth nor a post-falsification era [13].

Information-Theoretical Analysis

With this in mind, we now collate background assumptions needed to explain qualitative differences between humans and present-day AI. Firstly, while it is often implicitly assumed that all relevant information processed by humans can be reduced to classical bits – which a Turing Machine can model, we postulate that there is a general *asymmetry* between the ability to create new information of the type x and the ability to understand that new information x. Secondly, we introduce an epistemic artefact that has not yet been in the focus of AI research so far, has been termed "explanatory blockchain"[5] and is abbreviated with EB in the following. Novel EBs are solutions to problems constructed by interweaving blocks of explanations via the application of glue operations respecting a

[4] As stated by Frederick, when an observation conflicts with a theory, it could be that the theory is false, but also that the observation statement is false. Obviously, it could even be that both are false. Also, due to the fundamental impossibility of justification, probabilistic schemes are not helpful. It is for this reason that an epistemic stance that is independent of such considerations is required – which is given by the epistemic aim to strive for *better* explanations instead of attempting to reach "truer", "less false" or "more probable" ones.

[5] The narrower concept of explanatory blockchains (EBs) [7] facilitates an extension beyond the vaguer term of "explanatory knowledge" [45] which was frequently utilized by Deutsch but is problematic since present-day language AI *is* able to generate outputs that are colloquially perceived as "new explanations". However, *no* language AI has been collectively agreed upon by scientists and philosophers to have been able to generate new EBs respecting a rigorous epistemology as illustrated in Figure 4.1.

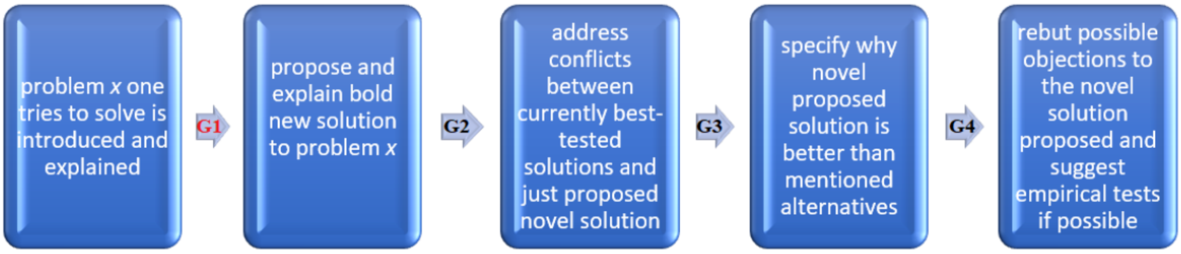

problem *x* one tries to solve is introduced and explained
G1
propose and explain bold new solution to problem *x*
G2
address conflicts between currently best-tested solutions and just proposed novel solution
G3
specify why novel proposed solution is better than mentioned alternatives
G4
rebut possible objections to the novel solution proposed and suggest empirical tests if possible

Figure 4.1: Exemplary epistemic total order for the generation of new EBs (the instructions are loosely inspired by an essay of Frederick [57]). Each glue operation x is indicated via a label G_x. EBs are a special form of explanatory information (EI) obtained by interweaving EI blocks via the step-by-step application of rational procedures sampled from a robust explanation-anchored, adversarial and trust-disentangled epistemology. Thereby, "trust-disentangled" signifies that the epistemic modus operandi is grounded in agreed upon criteria for *better* EBs i.e. it is orthogonal to any trust relation between involved entities – which means a better EB must be formulated such that metaphorically speaking it appears to defend itself against adversarial candidate EBs. In science, the specification of (direct or indirect) empirical tests in G_4 is the default condition.

rigorous epistemology as illustrated in Figure 4.1. In a general uttermost abstract way, we distinguish between two substrate-independent types of entities. While there are many ways in which one could formulate the distinction, here is a simple pragmatic one. Type II entities are all entities for which it is possible to understand linguistic explanations. From a linguistic perspective, the only *species* on Earth that would qualify as Type II is humanity. There may or may not be Type II entities elsewhere in the universe. Type I entities are all entities for which it is impossible to understand explanations. We postulate that all present-day systems that are commonly referred to as "AI" are Type *I* entities. Beyond that, while it holds that 1) Type I AIs can in theory forge the creation of any new *non-EB-like* information including texts perceived by humans as "novel explanations", it holds that 2) due to a gap of understanding, it is *impossible* for all Type I entities (thus also for those present-day "AIs") to reliably create *new* yet unknown EBs respecting an epistemic total order stemming from a rigorous epistemology as e.g. exemplified in Figure 4.1.

The last paragraph motivated why on our account, there must be a qualitative difference between present-day "AI" and humans. For a first glimpse on why new EBs could be epistemically special, it may be expedient to consider that their format corresponds to (or can be easily converted to) the format that was underlying all of humanity's best tested scientific theories (including formulations of laws of nature), best patent applications and best philosphical frameworks – at a time when those were new. Indeed, the lucrativity of intellectual property theft may be linked to the unestimable value of new EBs, generic new word chains that are so strong that their generation procedure cannot be forged. Our statement that it is impossible for present-day AI to create new EBs is a falsificable scientific statement. It can be made problematic by experiment but has not yet been falsified – despite growing hypes and overestimations of AI capabilities. Overall, the creation of new EBs may *not* be imitable because it cannot be predicted by scientific means. As stated by Popper, it is impossible to scientifically predict the future of knowledge creation [103]. On the whole, one could consider new EBs as representing the best form of recipes for novel unpredictable affordances that living entities can create. Moreover, given the definition of EBs, it now becomes possible to refine the description of our epistemic aim – be it in science or in philosophy. Namely, one can now specify that *our epistemic aim can be to achieve better and better new EBs.* Beyond that, we stress that we do *not* see the difference between Type I and Type II entities as a matter of degree. Instead, we conjecture a multi-level qualitative difference. We suspect that to understand EBs involves a nested understanding of all lower-level types of information of which no step can be skipped. On that account, while it is in theory *not* physically impossible to achieve a Type II AI *from scratch* as there is no law of nature that forbids it, we suspect it to be as hard as the task of constructing a novel universe e.g. by manufacturing black holes [117] – not physically impossible but practically so challenging that we believe that nowadays, there is exists no research on this planet able to achieve it in practice[6]. But note that for the pragmatic context of this specific chapter focusing on the enhancement of epistemic security for responsible AI design and not on physical theories, the falsifiable claim that the creation of new EBs is impossible for all present-day so-called "AI" systems can be considered independently from the reason of why this could be the case. Thus, we only briefly touch upon the latter and remark that further integrating research is required.

In recent years, a common theme from multiple research contexts that could be indirectly or directly harnessed to explain the special epistemic status of new EBs is the idea that the whole is more than the sum of its parts or that life cannot be reduced to the constraints of a Turing Machine. This also includes new frameworks that refute reductionist core assumptions and that could be grouped in three main categories: 1) cosmological [39, 38], 2) mathematical [80, 81] and 3) superinformation-related [1, 6, 52] perspectives. When integrating these different angles, one could postulate that new EBs represent a special form of information in nature. The latter could obviously be falsified by a Type I AI facilitating a shortcut to the creation of new EBs. In the future, it may thus be of interest not only for AI research but also for physics-related areas and even for psychology and neuroscience to perform experiments testing the information-theoretical nature of new EBs. Finally, one might ask whether one can use new EBs to test for Type-II-ness. The answer is yes and no. Indeed, it is possible to design a test but due to it being dependent on whether a Type II entity is *willing* to participate or not, blind tests for Type-II-ness could only be of *asymmetric* nature – next to moreover being of substrate-independent

[6] The latter may *not* apply to a reasoning about cases where instead of starting from scratch, one would e.g. consider suitable conscious biological Type I entities that already embody almost all required epistemic components except Type II language [7].

nature. An example for such a test framework is the Type-I-falsification-event-test [7] with the following asymmetric outcomes: while positive results can be mapped to a *homogeneous* group of entities of not nearer specified substrate whose Type-II-ness has been corroborated via new EBs they generated, negative results are ambiguous since potentially *heterogeneous* in that the test subject of not nearer specified substrate could be a Type I entity or it could be a Type *II* entity that was not willing to participate, not yet ready, not interested in the topic and so forth.

Theoretical Candidate Solutions

Having said that, we come back to the topic of epistemically-sensitive AI design. Given the framework introduced in this section, we conclude as follows. Firstly, seen from a cybernetic [19] angle, one could state that for requisite variety, we must account for the danger of Type I AI's theoretically permissible ability to forge any information as long as it is non-EB-like. Indeed, as the practical risk instantiations described in Section 4.2.1 already adumbrated, this asymmetry between the ability to create information of the form x and to understand that information x could lead to various epistemic threats if not explicitly anticipated. Secondly, realizing that new EBs are the only form of information that we understand which cannot be forged, we must try to focus our epistemic efforts at that level and ideally use Type I AI to augment our EB creation abilities. Thirdly, when combining the two last insights, it becomes apparent why a solution to deepfake science [7, 14] can be based on new EBs and has the freedom to focus entirely on the contents of new submissions instead of the sources and their substrates. (In a nutshell, the question should *not* be on whether a given contribution has been generated by present-day AI or by a human. Instead, a better question for scientists is on whether the contribution encodes a better new scientific EB in comparison to the EBs that are already available. Since Type I AI can only forge novel *non*-EB-like information and it holds more generally that new EBs cannot be forged [7], scientists have to invest cognitive resources to generate those and *cannot* craft an effortless pipeline to cheat. Thus, the reliance of science on new EBs allows the possibility of staying epistemically shielded [14] from any forgery.) Fourthly, when considering the idea that Type I AI could in theory be utilized to imitate everything that is imitable, it seems absolutely advisable to avoid the deployment of any robotic Type I artificial general imitator in conventional real-world environments as the latter could in practice risk to appear indistinguishable from any Type II entity that does not actively decide to participate in the creation of novel EBs given a specific context. Fifthly, it is easily conceivable that an artificial general imitator in virtual reality (VR) may be easier to simulate. While this may point at VR security issues to consider proactively, it may simultaneously offer a suitable counterfactual testbed for epistemically-sensitive AI design.

4.3 Practical Recommendations

4.3.1 Cyborgnetic Creativity Augmentation

In Section 4.2, we explained why on theoretical grounds, Type I AI could reliably forge the creation of any *non*-EB-like information – despite the described fundamental underlying limitations when it comes to understanding. This means there may be no theoretical limit on the accuracy of non-EB-like forgery using Type I AI. This may explain why AI did not only achieve predominance in game settings such as Go, but successes could also extend to the generation of new patterns helpful in drug discovery [72] up to the generation of novel texts perceived by humans as explanations encoding political propaganda [123]. In this light, it became apparent that new EBs – which cannot be forged (neither by Type I nor by Type II entities [7]) – are of uttermost epistemic importance. Interestingly, Peirce stated that signs are the only entities with which we can have a transaction [122]. In the light of the aforesaid, one could state that in constrained settings, it is only with new EBs that we can have Type-II-only transactions. However, it depends on the will of Type II entities on whether they decide to engage in the creation of new EBs. In the long-term, in blind contexts where this is not habitually implemented, Type I AI can induce epistemic threats since non-EB-like forgery can lead to situations where humans and present-day AI would become indistinguishable. Hence, it seems important to *empower* Type II entities such as humans to have the practical *option* to always create novel EBs when required and if desired. In short, one novel research avenue for a responsible, epistemically-sensitive AI design would be to specifically craft Type I AI *that augments people in EB creation processes.*

In this vein, the new meta-discipline of cyborgnetics [7] concerned with the mitigation of socio-psycho-techno-physical harm suggested to harness language AI for a targeted *cyborgnetic creativity augmentation* in crucial EB-based tasks such as threat modelling. A *cyborgnet* is a dynamic context-dependent functional template that can be described by a directed graph composed of *at least* one Type *II* entity and one Type *I* entity as introduced in Section 4.2.2. A *cyborgnet* is a highly generic term and is not to be confused with the much more narrow concept of a cyborg. Since cyborgnetics generically regards tools including language as a form of technology, the first language-cognizant humans already instantiated a cyborgnet. Thus, both an individual early human in the stone age and a modern cyborg equipped with an eyeborg such as Neil Harbisson [78] are an example of a cyborgnet. Moreover, while an entire security team of human researchers can act as one cyborgnet, it is possible to encounter hierarchies of cyborgnet networks including complex nested variants. Also, within a cyborgnet, relations are *not necessarily* bidirectional. In this way, the cyborgnet concept can also account for special cases such as e.g. monologues or safety relevant cases such as a guard convinced to have perceived a person which is later deconstructed to have been an unconscious misjudgement based on affective realism [61].

Cyborgnetic creativity augmentation using language AI could be of interest for threat modelling in cybersecurity, AI safety and security [7, 13] and more generally in any other subfield engaging in counterfactual risk analyses [131]. The key idea is that language AI trained on historical samples that are relevant for the risk domain in question can be utilized to *"create new ways to exploit hidden problems"* [63]. We emphasize that the just specified quote stems verbatim from the language model GPT-2 and has been inserted intentionally by us to serve as self-similar exemplification. In brief, the point is that one can utilize the obviously non-EB-like but potentially EB-creation-*stimulating* deepfake text that such language AI is able to generate to metaphorically speaking *look around corners*. Importantly, since the future of EB creation is unpredictable on theoretical grounds, language AI *cannot* serve as oracle tool. Indeed, any reliable oracle tool for the future of Type II entities must be impossible already merely due the possibility for EB creation. Thus, while there may be a few superficial similarities with ideas explored in the past such as the German project Cassandra [98, 121], cyborgnetic creativity augmentation serves a conceptually different purpose. The goal is *not* to predict the future. Instead, the goal is to contemplate plausible past *downward counterfactuals*[7] and try to act against them becoming our future by projecting plausible fictive better alternatives. In cyborgnetics, prior to the design of solutions for risk instances that occurred in the immediate past and that are documented in a "retrospective descriptive analysis" (RDA), one performs a "retrospective counterfactual risk analysis" (RCRA) which projects downward counterfactuals to the immediate counterfactual past. In this way, an RCRA adds breadth, depth and context-sensitivity to the space of RDA problem clusters – potentially leading to a formulation of better candidate solutions. This broad set of candidate solutions forms a "future-oriented counterfactual defense analysis" (FCDA) and consists of fictive but plausible *upward counterfactuals*[8] that are typically projected to the immediate counterfactual future. Language AI could be designed to support analysts in both RCRA and FCDA facilitating a multiversal approach [14, 110] to risk analysis. Finally, in the special case of risk analyses applied to epistemic security itself, it is clear that EB creation could profit from looking around corners and propagating through mental barriers by contemplating counterfactuals. An epistemically-sensitive AI design could then help to instantiate a *multiversal epistemic security* paradigm via the same language-AI-based cyborgnetic creativity augmentation mechanisms discussed.

[7] Ways in which an event *could* have plausibly turned out *worse* but did *not*.

[8] Ways in which an event *could* have plausibly turned out *better* but did *not*.

4.3.2 COOCA-Loop Meta-Paradigm for High-Risk AI Contexts

Given the theoretical background from Section 4.2.2 stating that Type I AI (i.e. including all present-day so-called intelligent systems) can neither understand EBs nor create new ones, one can anticipate a *comprehension bottleneck* that could arise in uninformed attempts to control it. For instance, one can start by examining the epistemic problems emerging in the extreme case of an intelligent system instantiating a classical OODA (Observe, Orient, Direct, Act) loop as end-to-end-Type-*I*-pipeline. In high-risk contexts and strategically complex domains, a reasoning via EBs may (and one could even state should) play a particularly important role. However, if the AI goal framework for the Type-I-OODA-loop pre-determined by humans would have been EB-based, the AI would not be able to enact its meaning in new contexts. The latter is given since it is considered to be impossible for a Type I AI to create new EBs. This represents a strong limitation to any conception of run-time "adaptivity" in EB-based decision-making including e.g. EB-based *moral reasons* [32]. A heterogeneous mixed scenario in which some functions are delegated to Type II entities but there exists a Type-*I-only* function does *not* solve the comprehension bottleneck problem as no novel EB-based message passing can be reliably implemented. Then, at first sight, consistent with the arguments presented in this paper it may seem recommendable to specify the requirement for high-risk contexts that *each single function* of an OODA-loop must be *cyborgnetic.* (A cyborgnet as a whole is always of Type II since it contains at least one Type II entity. Crucially, note also that a cyborgnet need *not* include any Type I AI since e.g. an individual human inherently lives in language and already fulfills the definition of a cyborgnet.) However, in the following paragraphs, we explain why strictly speaking, for epistemic reasons, one would then need to extend beyond the notion of an OODA-loop.

An OODA-loop could *not* epistemically be cyborgnetic because no reasoning in Type II entities begins by induction. In short, strictly speaking, no conscious OODA-loop actually starts with an observation. Instead, as already hinted by Popper [104], there must first be *a point of view* from which we actively sample the world – by what perception is inherently conjectural i.e. theory-laden. For this reason, a cyborgnetic OODA-loop would only stay an oxymoron. Thus, a first step is to explicitly add the following function: Conjecture (abbreviated with C in the following). As a second step, we explain why it is sensical to transform the Decide (D) function into a novel Co-create (C) function. Classically, in the AI field, decision-making is associated with a known set of options from which one has to choose. However, due to their own creativity capabilities and conscious choices, Type II entities can decide to create new options or even to destroy old ones. In brief, the space of options is strongly dependent on Type II creativity since it can ultimately contain the creation of new EBs (which can even include a revaluation of values [44]) for which it is impossible to reliably predict them ahead of time. Even where humans pre-specified an intention to throw dices, *"uncertain humans equipped with some dice at the time of* *moral decision making could throw that dice but could also unexpectedly (co-)create novel as yet unknown solutions on how to solve the problem"* [14] – something present-day "AIs" cannot.

As a last third step, one can now integrate the generic concept of AI-based cyborgnetic creativity augmentation exemplified in Section 4.3.1. In theory, this now becomes possible at the level of *each individual function* since each one is itself cyborgnetic. In general, to omit opportunities for creativity augmentation where adversaries practice it could be especially detrimental. It thus seems recommendable to implement it where practically feasible. To sum up, we just explained why for epistemic reasons, one requires the novel *meta*-paradigm of a cyborgnetic COOCA (Conjecture, Observe, Orient, Co-create, Act) loop for responsible AI design. Strikingly, some past approaches to responsible AI design are already intrinsically compatible with the *generic* COOCA-loop meta-paradigm and appear epistemically permissible as follows:

- ***Inter*-function-level:** Since *each single function* must be cyborgnetic, there must be at least one Type II entity in *each* function for EB-based communication *between* the functions. This is instantiated by some *human-in-the-loop* approaches. Also, recall that a Type I AI in a function is *not* obligatory.

- ***Intra*-function-level:** While each high-level function must be cyborgnetic, there is room for improvement *within* an individual function. There, *where feasible*, one can improve speed, scale and scope by harnessing *local* Type-I-OODA loops. This allows any of the three paradigms *locally within* the cyborgnet: human-before-the-loop, unsupervised loop and human-in-the-loop.

4.4 Conclusion and Future Work

In this paper, we explained why the evolving AI-based threat landscape in the deepfake era forces us to integrate *epistemic security* considerations in responsible AI design practices. Our *transdisciplinary* analysis used knowledge from diverse subfields including cybersecurity-oriented AI safety, adversarial AI, epistemological philosophy and cybernetics combined with new information-theoretical considerations to offer a rigorous theoretical basis for an *epistemically-sensitive* AI design – yielding the COOCA-loop meta-paradigm. We explicitly documented *practical* and then formulated *information-theoretical* limitations of present-day systems referred to as "AI". We explained that it is impossible for those Type I AIs to reliably create new explanatory blockchains (EBs). The latter cautions us not to *over*estimate those systems. However, we also examined the epistemic threats linked to malicious deepfake design including specifically the new frontier of *deepfake science* attacks. We explained why strictly speaking, in the long-term, there would be no theoretical limit to the accuracy with which Type I AI could forge the creation of any form of new *non*-EB-like information. We elucidated that while this in turn cautions us against *under*estimating yet underexplored AI facets that could be instrumentalized by malicious actors, we must simultaneously harness the opportunities for *cyborgnetic creativity augmentation* that it offers – also in order to defend against those threats. We gave an example for a special form of epistemically-sensitive AI design implemented as *multiversal* epistemic security paradigm via language-AI-stimulated EB creation.

Concerning the deepfake science threat, we explained that as long as science stays anchored in the creation of new EBs, it may stay epistemically shielded because those epistemic artefacts seem to be special and fundamentally inaccessible to any form of forgery including end-to-end-Type-I-pipelines. However, given our own epistemic grounding, it holds that we cannot know whether our assumptions are true. We only conjecture that they can be mapped to the best new EB that we have at present on that topic. As with all theories, it could be that the underlying assumptions will be falsified by experiment and be (provisionally) refuted by a better future EB. Our assumptions can be made problematic via the implementation of a Type I AI able to reliably create new EBs and be (provisionally) refuted by a novel theory able to i.a. provide a detailed explanation on how that Type I AI functions and why it is able to violate those. However, one may need to deepen the following two lines of thought in future work: 1) if our assumptions are refuted, this paper could have been generated by a Type I AI that did *not* understand it and 2) it would then be practically feasible to automate the creation of any new EB – and thus to automate science. In our view, this would represent an existential risk that would surpass any prior dual-use consideration. To put it plainly, while AI-powered drug discovery could *also* be used by humans to create biochemical weapons [125] and nuclear technology *also* allowed humans to destroy entire cities, an automated Type I deepfake science generator able to reliably generate *any* new EB could *also* facilitate the human-orchestrated destruction of... <generically fill in the blank>. Given our current best EB we assess that the latter is *impossible*, but is knowledge not fallible?

Chapter 5

The COOCA-Loop Solution – A Detailed Account

5.1 The Practical Problem: Can One Meaningfully Control The Type I OODA Loops of Present-Day Intelligent Systems?

How to *meaningfully* control present-day intelligent systems [18, 130, 134] became a topic of international interest in the scientific community and beyond. Many solutions (that will not be reviewed here) have been suggested ranging from human-in-the-loop [75, 137] to human-before-the-loop [16] approaches. In the next Section 5.2, I take a new cyborgnetic perspective and first deconstruct the adverb “meaningfully” using *substrate-independent* ontological distinctions related to cyborgnetic information types. Alongside, I extend the impossibility theorems of cyborgnetics [3] (the ITCs) to in total seven elements. I explain why Type I OODA loops are not a shortcut to a genuine “value alignment” with present-day Type I AI – which can in any case not be achieved on theoretical grounds as already reflected in the AI safety paradox [4]. Then, Section 5.3 delves into the practical implications for the *control* of Type I AI against the backdrop of the foregoing theoretical analysis. I explain why not only complementary *adversarial* simulations are required but especially why one needs to integrate the Type I OODA loops of present-day intelligent systems into *cyborgnetic feedback-loops* based on explanatory blockchains (EBs). Section 5.4 explains *emergence* phenomena paired with EBs from a *cosmological*, *mathematical* and *information-theoretic* standpoint.

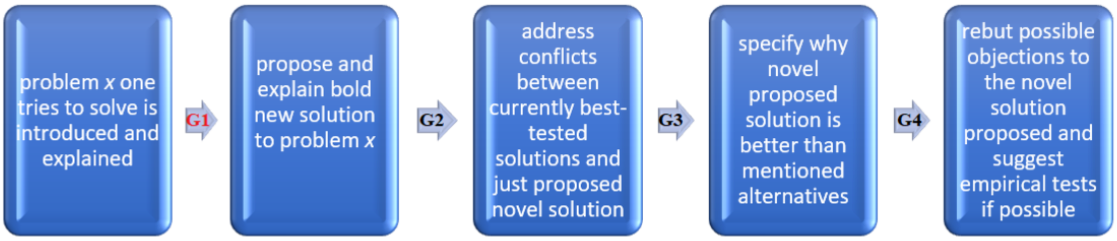

problem *x* one tries to solve is introduced and explained
G1
propose and explain bold new solution to problem *x*
G2
address conflicts between currently best-tested solutions and just proposed novel solution
G3
specify why novel proposed solution is better than mentioned alternatives
G4
rebut possible objections to the novel solution proposed and suggest empirical tests if possible

Figure 5.1: Exemplary *epistemic total order* for the generation of new EBs (the instructions are loosely inspired by an essay by Frederick [57] on how to write better philosophical papers). Each glue operation x is indicated via a label G_x. EBs are a special form of EI obtained by interweaving EI blocks via the step-by-step application of rational procedures sampled from a robust explanation-anchored, adversarial and trust-disentangled epistemology. Thereby, "trust-disentangled" signifies that the epistemic modus operandi is grounded in agreed upon criteria for *better* EBs (i.e. it is orthogonal to any trust relation between involved entities – which means a better EB must be formulated such that metaphorically speaking it appears to defend itself against adversarial candidate EBs). Examples for such more widely accepted criteria in science are for instance: a preference for theories that provide more novel falsifiable predictions than rival ones, theories that are simpler, more interesting or more aesthetically appealing.

5.2 Theoretical Answers

In this section, I first examine the intrinsically relational conception of "meaning" in the context of Type I AI control. I consider human morality to be mostly "explanatory" (but here in this case *not* necessarily based on explanatory blockchains (EBs)) by virtue of mostly apparently being linked to norms and values that are considered to be reasonable i.e. more precisely it involves either non-EB-like or ideally EB-like explanatory information (EI). For an exemplary step-by-step procedure to craft an EB, see Figure 5.1. However, as described in cyborgnetics [7], there is an *asymmetry* between the ability to *create* a specific form of information and the ability to *understand* that information. In this vein, since present-day AI does *not* understand EI (and by extension also *not* EBs) despite its ability to create non-EB-like EI, it is clear that once heuristic human moral models are encoded in it for purposes of control, they are not enacted in any EI space – even if the counterfactual branches that such AI could output for "explainability" purposes could be formatted as EI. The latter needs more consideration since of relevance once intelligent systems equipped e.g. with ethical goal functions [16] would be deployed in real world environments.

5.2.1 Meaning and Information in Cyborgnets

The Ѐdishe-theorem [7] is composed of multiple parts one of which explicitly mentions the concept of *shared* indexical and iconic information (SIII). Thereby, SIII specifically refers to indexical and iconic information that is shared in the common ecological niche of given animals e.g. simply during habitual collective activities (see also [22] for an in-depth analysis of semiotic details). More precisely, the Ѐdishe-theorem implies that while conscious Type I animals – mainly vertebrates, cephalopods and arthropods [23, 94] – are able to understand SIII, present-day AIs (corresponding to non-conscious Type I entities) are *not*. In the new Gatejeli-theorem that I present in the following, I introduce the notion of *collective* biological information (CBI) to refer to indexical information that is *collectively* shared in the ecological milieu of given living entities e.g. while currently occupying physically adjacent spots. Indeed, one can state that the exchange of physiological signals between organisms is a widespread phenomenon in biological milieus. Crucially, it is vital to note that in contrast to SIII, CBI does *not* presuppose consciousness. Then, the Gatejeli-theorem states that it is *impossible* for *non-living* Type I entities to *understand* CBI. Importantly, the theorem does *not* touch upon the ability to *create* CBI. For instance, nowadays it is cogitable that it might already be technically feasible to build a non-conscious and non-living Type I AI able to create new CBI, SIII and EI. However, that AI would still not be able to understand CBI. By contrast, I conjecture that the biological programs instantiated in many clearly *non-conscious* [94] but living biological Type I entities such as plants, fungi and bacterial biofilms are able to not only create CBI but also to *understand* CBI. Note that since the ontological distinctions in cyborgnetics are *substrate-independent*, in case extraterrestrial life would exist, it *could* obviously instantiate CBI understanding too. Generically, life seems to involve a dynamic physiological coupling of constant interactions with the physical environment [127] which can often comprise environmental stressors and other life forms.

For instance, within bacterial biofilm communities [53], communication is possible via ion channels [106]. Plants differentially respond to environmental stimuli as a function of whether they are nondamaging (e.g. touch, cooling, light) or destructive (e.g. wounding, burning injury) [94]. The former leads to specific action potentials while the latter causes differently characterized, slow wave potentials associated with defense responses to stress [94]. Fungi are able to form *"large networks on the forest floor that are too large to distribute nutrients through diffusion alone"* [60]. A *bidirectional* transfer of small RNA between plants and fungi is possible [129]. That being said, the *substrate-independent* view in cyborgnetics also implies that *artificial life must be possible.* In this vein, recent research on programmable, *functionally* designed *xenobots* [21] implemented on the basis of frog cells offer a first glimpse on what living (but still *non-conscious*) Type I AI could signify. These xenobots which have been described to emerge from cellular self-organization and whose "lifespan" (from days to weeks [85]) can be extended with nutrient-rich media [24] are e.g. able to *"navigate aqueous environments in diverse ways, heal after damage, and show emergent group behaviors"* [24]. Moreover, researchers recently analyzed the novel emergence of an unseen kinematic self-reproduction mechanism exhibited by xenobots [86]. In light of the ITCs, I already postulated earlier [8] that such non-conscious but living Type I xenobots could *reliably* instantiate an understanding of CBI. It is even conceivable that a complete automation of their design [85] (building on suitable cells) becomes practicable. Although an in-depth analysis is beyond the scope of this paper, the case of considering xenobots as artificial life has been elucidated elsewhere [36, 50, 86]. What is relevant for this paper is that next to *not* being able to understand SIII and EI as implied by the Èdishe-theorem [7], the Gatejeli-theorem implies that in addition, the however advanced present-day non-living intelligent systems are *not* able to understand CBI – while even xenobots made of cells may realize the latter.

5.2.2 The Cyborgnetic Ladder of Understanding

Before coming back to the control of Type I intelligent systems, this section first introduces other intermediate levels of information that must be introduced. Firstly, one could consider something that one could call molecular and other, ionic information (MoI) – a subtype of which could be organic molecular information (OMI). Both MoI and OMI could be perhaps described to entail a total order on their subcomponents. For instance, research utilizing methods from computational linguistics corroborated *analogies* between language and *organic* chemistry [29]. I suspect these analogies to simply pertain to the existence of an underlying well-formed sequential ordering. Hence, I assume it to represent a structural and not content-related commonality. In this vein, it may thus not be suprising to notice that non-conscious and non-living Type I AI (including natural language processing models) may be able to create new MoI as allegedly corroborated recently in the context of drug discovery [43, 72, 99, 126, 140]. However, when running in silico on Earth, this non-living Type I AI does *not* instantiate an *understanding* of this new MoI. By contrast, biological *cells* can navigate complex routes via "self-generated chemotaxis" [124], i.e. by creating own local chemical attractant gradients. In this way, these cells were able to persist and travel over long distances through *new* complex microfluidic mazes (which would not have been possible with simple chemotaxis [124]). I assume that the biochemical enactment of such living entities instantiates OMI understanding[1]. Beyond that, it is also thinkable that *in the biosphere on Earth*, any CBI understanding already intrinsically implies an OMI understanding – which may be one of the reasons why the potentially CBI-cognizant xenobots [21] could persist for multiple days given that they were fabricated on the basis of frog *cells*. On the whole, it may appear reasonable to

[1] As another example, in the daily life of an eukaryotic cell, chemosensing or tracking of other cells is part of the habitual set of excitable actions [127].

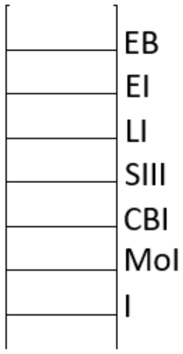


Figure 5.2: Simplified illustration for *the cyborgnetic ladder of understanding.* Following cyborgnetics and cynet information theory [8], there exists an *asymmetry* between the ability to create information of the form x and the ability to understand x. In theory, for all steps x on the ladder *except* the last step of EBs, it is possible to create x *without* understanding x. For the special case of EBs, it holds that only Type *II* entities (of which humans are an example) are able to understand EBs *and* it is *only* Type II entities that are able to create *new* – i.e. previously unknown and non-plagiaristic – EBs. The latter could be falsified by experimentally demonstrating a Type I AI able to *reliably* create *new* EBs and it could be provisionally refuted by additionally explaining how it was programmed. Note however that a refutation of the cyborgnetic ladder would signify that all science could be automated (a potential existential risk for humanity) and that e.g. cyborgnetics and the cyborgnetic ladder itself could have been invented by a Type I AI that did *not* understand it.

consider MoI as an intermediate between binary information (simply abbreviated with "I" in the following) and CBI. Though, I currently suspect CBI and MoI to be so closely interlinked that once MoI understanding became possible in the terrestrial biosphere via the emergence of cells, at least *the potential* for CBI understanding was given too even if it might have only manifested itself once other MoI-cognizant living entities arrived in the physical vicinity of such cells. Interestingly, in vitro studies corroborated that *collective* cell dynamics in *closed* environments were able to emerge merely on the basis of the behavior of *single* cells *"through a sustained memory of cell polarity"* [70]. In sum, while it is plausible that MoI understanding emerged first, it also seems plausible that the potential for CBI understanding was already present at that point. Thereby, it is clear that CBI can extend to much wider repertoires of collective behavior than the first MoI-cognizant creatures may have been able to – as e.g. corroborated in the development of eukaryotic cells [127]. In a highly *simplified* way, since CBI appears to simultaneously also be MoI but not the converse, it can be seen as a proper subset of MoI and a distinction makes sense.

Secondly, I briefly introduce a straightforward additional notion that is likewise not understood by present-day intelligent systems: linguistic information (LI). Here, LI can be positioned between SIII and EI. While EI was very narrowly linked to statements about the what, how *and* importantly *why*, LI represents a more general notion. More precisely, in cyborgnetics, an LI medium can be defined as an information medium[2] with the additional properties that: 1) its attributes are symbols, 2) its set of attributes has a total order relation $\preceq$ defined by a Type II language. In short, LI does not only include linguistic statements pertaining i.a. to the "why" but simply refers to *all* linguistic statements within a Type II language. (Interestingly, in language, symbols can additionally function as both icons and indexes [22]. Generally, some icons can act as indexes too.) Before discussing the implications for the control of present-day intelligent systems, one could now retrospectively contemplate the information types displayed in the Èdishe-theorem and conjecture the following *highly simplified* chain of nested relations via proper subsets: $EB \subset EI \subset LI \subset SIII \subset CBI \subset MoI \subset I$. Henceforth, I denote the visually *reversed* form of this chain-like postulate, the *cyborgnetic ladder of understanding* [8]. I conjecture this cyborgnetic ladder illustrated in Figure 5.2 to be metaphorically isomorphic to an epistemic artefact of *understanding* where all steps are *obligatory.* To put it very simply, I assume that with $1 \leq x \leq 6$ and $x \in \mathbb{N}$, to *understand* a step $x + 1$ on the cyborgnetic ladder, it is impossible to skip the previous step x. I call this just mentioned new impossibility theorem, the *Talièshe-theorem.* With other words, *it is impossible to skip a step on the cyborgnetic ladder if the goal is to understand the next one*[3], i.e. a shortcut for understanding is impossible.

5.2.3 Shortcuts to the Control of Type I Intelligent Systems?

Now coming back to the topic of present-day intelligent systems equipped with ethical goal functions (EGFs), three cases have to be analyzed: 1) non-EI-like but LI-based, 2) non-EB-like but EI-based and 3) EB-based EGFs. Namely, as background assumption, it seems plausible that, while human morality mostly involves non-EB-like EI or EBs, it *at least* harnesses LI which is on step 5 of the cyborgnetic ladder. Consequently, it becomes clear why "value alignment" with present-day intelligent systems is impossible. For it to work, the system would have to *at least* understand SIII (step 4). However, apart from trivially Type II entities, the only entities known so far whose physical substrates are able to understand SIII, are *conscious* Type I animals. At the same time, according to the Èdishe-theorem [7], it is impossible for conscious Type I animals to reliably understand EI. At that point, it becomes straightforward to extend the Èdishe-theorem to LI by

[2] Here, information is necessarily instantiated in a *physical* substrate since cyborgnetics borrows the concept of information from constructor theory of information [48] where information is grounded in physics and *not* merely floating in an abstract mathematical sphere.

[3] For instance, in order to understand EI (step 6), one must first understand non-EI-like LI (step 5). In order to understand an EB (step 7), one must first understand non-EB-like EI (step 6) and so forth.

adding that it is impossible for conscious Type I animals to understand LI by virtue of it being encoded in a Type II language as depicted earlier. In sum, *a Type I intelligent system could not even meet the condition for step 5 (LI) of the cyborgnetic ladder*, since it already takes Type-II-ness to reach this step.

The latter could be e.g. falsified by implementing and explaining a Type I intelligent system instantiating: 1) a shortcut to SIII-understanding *without* Type I or Type II consciousness, 2) a shortcut to LI-understanding *without* Type *II* consciousness, 3) a shortcut to EI-understanding *without* Type *II* consciousness and/or 4) a shortcut to EB-understanding *without* Type *II* consciousness. Depending on the cases, it could make the Èdishe-theorem and the Talièshe-theorem but also the Adije-theorem [7] highly problematic. As long as the latter have not been refuted, one can recapitulate with the statement that the comprehension gap between Type I intelligent systems and human moral models encoded e.g. in the form of EGFs can *not* be bridged. Upon closer analysis, it seems as if a *vanilla* EGF-based socio-technological feedback-loop with present-day Type I intelligent systems [16] simply stays a purely bit-based loop (a non-MoI-like information (I) loop, or in short a non-MoI-like I-loop) – the only kind of message that is reliably communicated throughout the active nodes of the underlying cyborgnet [7]. Especially, even if humans would have utilized EB-based EGFs, it is instead non-MoI-like bit streams that are enacted in a non-MoI-like I-loop. This signifies that instead of EBs, it is non-EB-like I that is governing the decision phase of the system's OODA-loop. The Type I intelligent system represents a *cyborgnetic comprehension bottleneck*[4]. Note that the same would also affect any conventional human-in-the-loop solution as it does *not* change the *ontological type* of the information that is transmitted from end-to-end. Given these insights, an old pertinent question may arise again: is the control of Type I intelligent systems deployed in real world environments impossible? I provide a new answer to this question and explain why a new *cyborgnetic detour* is required.

5.3 Practical Implications of Theoretical Answers

In light of the cyborgnetic comprehension bottleneck, one could try to control Type I intelligent systems by integrating those in an ideally EB-based cyborgnetic OODA loop i.e. where the *decide* part is not purely bit-based, but grounded in EBs. However, note also the fact that for security reasons, neither the *observe*, nor the *orient* nor the *act* part of the loop can be *entirely* delegated to Type I AI due to the vulnerability to maliciously crafted adversarial examples [135] or also simply to many out-of-distribution events –

[4] This is interesting since in AI safety, *the reverse* is often assumed. Namely, that humans may represent bottlenecks to superintelligent AI systems in integration scenarios due to the restricted cognitive capacities of the former.

which I postulate both simply represent other instantiations of the cyborgnetic comprehension bottleneck. One may now ask the question on what the use of a deployed Type I "intelligent system" could be at all if it can neither safely autonomously observe, orient, decide nor act without support from a Type II entity – be it at the decision level or even as additional sensor and actuator to avoid catastrophic failures. In my opinion, the answer requires a cyborgnetic detour that reformulates the task of interest. Instead of serving as OODA-substitute, intelligent systems would ideally be harnessed for cyborgnetic creativity augmentation [7] enhancing old functions and additionally complementing a novel kind of function that I label as *co-create* (C). However, one has to consider the following three subtle details. Firstly, the C function is substrate-independent and even entity-independent since entirely *content-based.* This signifies nowadays that both humans and the AI systems would perform the C function. Secondly, humans *and* the Type I intelligent systems would analogously jointly engage in *observe*, *orient* and *act.* Thirdly, strictly speaking, from an epistemic perspective, no *conscious* OODA-loop actually starts with an observation. That would correspond to a pure bottom-up induction which is impossible according to Popper [104] since *a point of view* (including a conjecture in the case of Type II entities) is always required in the first place.

Thus, I now conjecture that Type I intelligent systems could ideally be utilized to *augment* EB-based COOCA-loops (Conjecture, Observe, Orient, Co-create, Act) with regard to each function. This cyborgnetic detour is *not* affected by the cyborgnetic comprehension bottleneck because each function is now performed by *both* Type I *and* Type II active nodes of the cyborgnet. Since the entire cyborgnet is now foregrounded and cyborgnets as a whole are of Type II, it becomes possible to have a message transmission communicating EBs. The utility of the Type I intelligent system now relies in systematically augmenting the human at each step. In this vein, I have recently explained how *cyborgnetic co-creation* sessions with both Type I and Type II entities could be implemented in social virtual reality where NPCs driven by language AI could verbally stimulate human creativity [7] in explanation-anchored co-creation design fictions, serious games or educational gamification. In the case of Type I intelligent systems deployed in real world environments, it could for instance make sense to design them as virtual or cyberphysical robots that are specialized in any of the five functions of the COOCA-loop. Thereby, note that when applied to slow time scales in cyborgnetics, one could map the *conjecture*, *observe* and *orient* parts to a retrospective descriptive analysis (RDA), the *co-create* part to both retrospective counterfactual risk analysis (RCRA) and future-oriented counterfactual defense analysis (FCDA) whilst the final *act* part could be associated with enacting the FCDA results.

For illustration, consider an EB-based cyborgnetic detour that would consist in implementing three different Type I intelligent systems: one for the RDA, one for the RCRA and one for the FCDA. Each one would assist humans in different ways – virtually but also physically where needed. For instance, a 3D avatar visually displayed on the desktop embodying a hybrid active RDA agent with textual inputs could ease the crawler-based collection of RDA samples online according to clusters pre-specified by humans. A subsequent immersive co-creation session in social virtual reality could involve an RCRA-and-FCDA agent taking the form of an avatar powered by a language model and providing speech inputs to stimulate the generation of novel RCRA and FCDA samples – importantly this would also contain *adversarial* considerations e.g. for adaptive attacks. Finally, a physically deployed FCDA agent with advanced motor capabilities could assist in repetitive and predictable tasks during a mission (such as e.g. simply carrying equipment) or another physically deployed FCDA agent could instead serve as sensor *complementing* human personnel in the detection of blind spots.

While such complementary Type I agents would internally still instantiate their own *local* non-EB-like I-loop, the information transmitted *between the individual functions* in such an EB-based cyborgnetic detour would be *EB-like* since every step is *jointly* involving Type II entities which are inherently able to pass the message. Note that it still appears reasonable to equip the local Type I intelligent systems with flexible *local* EGFs constraining the state-action space with cognitive-affective criteria that could enhance *creativity* but also with parameters encoding *moral conceptions* – even if only enacted at the level of non-MoI-like I-loops. Such EGFs could be crafted by indvidual users in specific contexts whereby designers could provide default settings according to own inclinations and recommendations. This task can be implemented with an input-agnostic scientific scaffold such as augmented utilitarianism [4] as described in-depth recently [15]. To recapitulate, the rationale for the cyborgnetic detour via a COOCA-loop is as follows: the communication process of interest to the cyborgnetician is between the five individual functions: conjecture, observe, orient, co-create and act. What is of relevance is to *only* allow EB-like streams of information *between* the functions. For it to be realizable nowadays, humans would need to be the ones sending and receiving messages between the functions. However, locally, *within* a single function, humans can integrate advanced Type I intelligent systems as local assistive agents with local EGFs to augment them in any suitable activity. Knowledge from scientific research including artificial creativity augmentation [11] can support a tailored design for these assistive agents.

5.4 Conclusion and Explanatory Contextualization

5.4.1 Preliminary Concluding Remarks

In this paper written for purposes of self-education and serving as ephemeral mental clipboard, I have asked the question of whether it is indeed possible to meaningfully control the Type I OODA loops of present-day intelligent systems. Firstly, I analyzed meaning from an information theoretic perspective and introduced the *Gatejeli-theorem* stating that it is impossible for non-living entities (which includes present-day intelligent systems) to understand collective indexical information – a simple epistemic artefact whose understanding is by contrast even already instantiated in bacterial biofilm communities. Secondly, connecting the former to the Èdishe-theorem elucidated elsewhere previously [7], I formulated the *Talièshe-theorem*. To put it very simply, the Talièshe-theorem states that, when it comes to *understanding*, it is impossible for any entity to skip a step when attempting to climb the *cyborgnetic ladder*. This cyborgnetic ladder comprises seven consecutively binding steps[5] if the goal is to *understand*: 1) information (I), 2) molecular and other, ionic information (MoI), 3) collective biological information (CBI), 4) shared indexical and iconic information (SIII), 5) linguistic information (LI), 6) explanatory information (EI) and finally 7) explanatory blockchains (EB). Then, since intelligent systems do not even exhibit CBI understanding, a *cyborgnetic comprehension bottleneck* emerges. In order to avoid human moral models to be caught in the non-MoI-like (and thus intrinsically non-EB-like) OODA-loops of intelligent systems, which would pose a serious threat to any ambitions to control those systems, I postulated that a systematic *cyborgnetic detour* is required.

While there is no shortcut that would make an EB-like Type I OODA-loop possible, the cyborgnetic detour consists in creating and implementing a different strategy with five functions: a COOCA-loop (Conjecture, Observe, Orient, Co-create, Act) representing a Type-II-and-Type-I-EB-co-creation endeavor [7]. Nowadays, the rationale would be that *within* a single function, humans can utilize intelligent systems as assistive agents augmenting them at any suitable activity, whilst the communication *between* each of the five functions is EB-based and solely conducted by humans. A straightforward objection would be that COOCA loops being of Type II would be often *much slower* than their Type I OODA counterpart. However, the price of security is eternal creativity [4]. Deliberate and spontaneous creativity are of Type II and are inherently slower than intelligence-focused optimizations in other specific tasks. Also, note that the speed of the *local* OODA loops situated *within* individual functions and governed by local ethical goal functions are not affected by the slow thinking modes and humans may in addition be able to perform faster on average in comparison to cases where a Type I assistive agent would be lacking. Nowadays, there is often the assumption that humans have to steadily increase the speed at which they operate. However, it is important to question the latter. We might need to take time for creativity, otherwise we could be swept away by the non-sensical OODA-loops of at-present not even MoI-cognizant Type I entities which instantiate a cyborgnetic

[5] When humans are born, their physical substrate already instantiates the first 4 steps and they are immersed in step 5 ab initio which concurrently moulds their Type II brains. Mostly, it does not take long until step 6 is reached [82]. However, while the potential is always available *covertly*, whether step 7 is *overtly* considered depends on each individual human.

comprehension bottleneck but are perceived as being "intelligent" and exhibiting comptency or even agency.

Upon recalling that the cyborgnetic ladder pertains to *understanding* (and *not* creating) knowledge, it can appear abstruse to apply it to humans. However, while brain-centered perspectives may obfuscate it, the claim is that the existence of any Type II entity, which includes humans, instantiates indeed the *entire* cyborgnetic ladder. In the following, I briefly depict how it applies to humans. Firstly, humans are able to *consciously* understand EB-like EI, non-EB-like EI and non-EI-like LI. Secondly, humans are also able to *consciously* understand non-LI-like SIII as instantiated in the iconic and indexical forms of communication that many humans are able to carry out with non-human conscious animals or with human infants in their collective enactment in continuously shared environments. Thirdly, it becomes important to note that neither non-SIII-like CBI, nor non-CBI-like MoI, nor non-MoI-like I understanding require any form of consciousness. The cells in the human body (including the cells in the brain) are themselves clearly *non-conscious* [94] but instantiate an understanding of non-SIII-like CBI and MoI, otherwise humans may not be able to survive without artificially having figured out how to use Type I technology to reliably allostatically regulate their existence. Fourthly, certain subparts of cells may not have to instantiate more than binary electrical non-MoI-like I understanding. However, beyond that, there is an example of *consciously* understanding non-MoI-like I in humans: their ability to reliably artificially implement contemporary computer hardware to perform calculations is an instantiation of non-MoI-like I understanding or in short, *understanding classical bits.* Perhaps, in this 21st century, it now becomes important not to get lost in the I-loops of our intra-cyborgnetic interactions with non-living Type I AI – trapped at the lowest *understanding* level of ourselves. This finally leads to the topic on emergence vs. reductionism.

5.4.2 EBs Explaining Emergence Phenomena of EBs

In the AI field, some assume that Type I AI could learn any thinkable task and that all tasks can be solved on the basis of bits. I refer to this prevailing stance as *the reductionist paradigm.* In diverse past cyborgnetic books [5, 8, 10], I elucidated multiple facets on *why* an EB is more than the sum of its parts. On the whole, by focusing on the task of creating *new EBs*, cyborgnetics and its idependent branch of cynet information theory [8] refute the reductionist paradigm. This refutation can be complemented by various novel developments in physics and beyond. Overall, some of the currently best available EBs on that subject can be e.g. classified in three categories: 1) cosmological lines of reasoning, 2) mathematically focused analyses and 3) superinformation-related hypotheses. I very briefly introduce some key take-aways from each category.

Cosmological Perspective

When considering biocosmology [40, 38]– a framework very recently introduced by multiple known physicists – it becomes clear that most present-day AI (which is non-living Type I AI) including so-called intelligent systems may not exhibit requisite variety when compared to Type I and Type II *life* due to the immense space of degrees of freedom that living entities add to the universe as a whole. Already the degrees of freedom exhibited by stars[6] may not yet be attained by even the most advanced present-day non-living Type I AI. In addition, from the perspective of cynet information theory and its independent cosmologically-focused branch, while the set of possible functions for Type *I* life may indeed tend to grow steadily [40], may reach infinities and is unpredictable ahead of time [79] as explained by biocosmology [40], it is important to *additionally* consider the special case of Type *II* life. For a compressed summary on how biocosmology can be extended by applying a cyborgnetic lens, see [9].

Indeed, Type II life, through the ability to consciously understand what a construct such as "possible functions" means, can consciously decide how to enact, enlarge but also to *reduce* those. Moroever, Type II life can also decide to behavioristically mimick selected trajectories and distributions. On the whole, Type II beings are *not* bound to biological imperatives or to the consideration of functions that are solely in the service of what is often described as "biological fitness". While biological entities can harness stochasticity at various levels [97, 96] leading to a partially sighted process including phenomena such as targeted mutations [97], Type II entities can craft EBs about what "stochasticity" signifies and can consciously employ it. Type II entities can literally even consciously manufacture selected mutations for their socio-culturally constructed goals. Moreover, voluntary suicide or the conscious destruction of the biosphere become possibilities. The result is that when trying to grasp something like the "number of possible functions" for Type II entities, one suddenly encounters aggregate abnumeral infinities – what Peirce called *supermultitudinous collection* [73].

Mathematical Perspective

Following Kauffman and Roli, the affordances that *living* entities enact in their biological milieu *cannot* be expressed via set theory [80]. They state that *"we can create no*

[6] Perhaps a hypothetical fictive future nuclear fusion reactor based on non-living Type I AI could reach this level. However, the application of non-living Type I AI to significantly improve nuclear fusion is currently only at the beginning with deep reinforcement learning [42]. Further, it is thinkable that active inference [41] could enhance the required non-living intelligent system – which may however come of the cost of predictability. What is more, also in this relevant context, one must consider context-dependent harm models such as augmented utilitarianism [4] and one must inject Type-II-ness for a cyborgnetic risk management applied to a COOCA-loop.

mathematical model of the diachronic evolution of the biosphere based on set theory" [80]. One cannot mathematically predict those *ahead of time* [81]. Concerning Type II life, I postulate that one must strictly speaking consider the mentioned concept of a supermultitudinous collection – which is neither a set nor even a category from category theory. It is an ultra-dense condensate of genuine infinity of which there exists no higher order. As described by Peirce, the elements of such a collection are *not* points, but triadically *interdependent potentials* (see [73] for an in-depth explanation). In brief *"a supermultitudinous collection sticks together by logical necessity. Its constituent individuals are no longer distinct and independent subjects. They have no existence [...] except in their relation to one another"* [73]. I assume that it is for this reason that no mathematical formula can predict or postdict the creation of an unknown new EB and no formula can cover the entire potential of cyborgneticity.

Superinformation-Related Perspective

Interestingly, Aerts and Beltran [1] recently corroborated that new EI in the form of stories (such as Winnie The Pooh) can be interpreted as a special form of superinformation[7] since they were able to experimentally corroborate that – when directly compared to the classical Maxwell-Boltzmann statistics – *Bose-Einstein-statistics* represented a *superior* model for those texts. The authors elucidate that the latter may represent an explanation for the appearance of Zipf's law [1] known in computational linguistics. Prior to that, in the second cyborgnetic book [8], I postulated that specific new EBs (for instance anagrammatically encrypted ones and generally those intermingled in non-EB-like EI) are expressible as a new form of superinformation [7]. (i.e. more than assembled I pieces). In the third one [5], I generalized it to the statement that any *new* EB is a form of Type *II socio-psycho-techno-physical* superinformation while *new* non-EI-like LI or *new* non-EB-like EI can act as Type *I* socio-psycho-techno-physical superinformation – since it can be forged by Type *I* entities even though those do *not* understand it. (Note that since new EB forgery is impossible [3], new EBs represent a stronger form of superinformation that is only accessible to Type *II* entities [7].) Finally, what is conventionally described as quantum information can be described as a special case of *physical* Type *I* superinformation[8].

[7] Superinformation is a *scale-independent* term introduced in constructor theory of information [48]. Quantum information is only one special possible form of superinformation.

[8] When considering quantum information, it is important to keep in mind that it involves *mathematical* and thus substrate-independent formalisms and one must thus avoid the substrate-dependency-fallacy of mentally *a priori* reducing it to miniscule e.g. subatomic or atomic scales. What seems relevant to superinformation instantiating a non-classical paradigm are the notions of *entanglement*, *superposition* and *encryption* [20] – all of which are *not* a priori tied to a specific scale. In modern days, the discipline of quantum biology [83] gained momentum [52]. While only in its infancy, it already provided some experimental corroborations of quantum effects at many corresponding steps of the ladder: for instance

5.4.3 Synopsis

Due to cyborgnetic *emergence* phenomena, a shortcut for *understanding* is impossible. Simultaneously, it holds that in theory, the *creation* of any information form *except the creation of new EBs* can be forged. This asymmetry between the ability to create information of the form x and to understand x leads to various fundamental issues. While present-day intelligent systems could be designed to output strings that are perceived by humans as representing EI, it is important to keep in mind that there is no underlying understanding by such Type *I* AIs in order to avoid honey mind traps [7]. A Type-*II*-infused *COOCA*-loop is recommended as cyborgnetic detour for meaningful EB-based intelligent system control.

5.4.4 Silicon vs. Carbon

Strictly speaking, OMI would *not necessarily* correspond to a substrate-*independent* ontological distinction that is valid on all planets that could exhibit life forms or at least valid for all *technically feasible* system designs. This is due to the hypothetical possibility of *silicon*-based life (with silicon as backbone for molecules carrying biological information) as alternative to carbon – for specific chemical reasons [100]. While this option is mostly discarded in astrobiology and does not count as focus in the exploration of extraterrestrial life [90], there would in theory be nothing *fundamental* that would prohibit Type II entities to intentionally engage in a plan of creating artifical silicon-based life which could then at some point instantiate CBI understanding. Since predominantly silicon-based molecules do *not* count as organic, this silicon-based path to CBI would have seemingly skipped the OMI requirement. However, it is crucial to note that even then, another suitable form of *MoI* understanding would still need to manifest itself – with life generally being closely linked to *chemically* modulated excitability [127].

Also, specifically on this planet Earth, silicon-based life could come with complications. For instance, it has been elucidated that a silicon-based organism inhaling oxygen would exhale silicon dioxide – reminescent of rocky sand [90]. Thus, despite the higher relative abundance of silicon on Earth in comparison to carbon (which is the inverse when seen at the level of the entire universe), it seems more promising to start with carbon as backbone for OMI – as is the case in all life on Earth down to cells [100]. On the whole, one can identify two reasons why the silicon-based path to life is not yet directly a promising outlook in light of the low understanding capabilities of present-day intelligent

at the level of DNA mutations [118], in cell-related oxidative stress mechanisms [108, 136], in living but non-conscious Type I entities such as plants or in conscious Type I life such as birds [89, 119, 132]. Human magnetosensitivity [33, 55, 113] (but so far *without* consciously accessible sense of it) and its conjectured link to spin dynamics [136] may thereby perhaps offer novel avenues for future yet unknown new EBs.

systems and all other present-day non-living Type I AIs. Firstly, since the present-day Type I AIs implemented on Earth are primarily instantiated on a silicon-based hardware *on Earth* but OMI understanding seems to belong to requisite variety for life at least in the terrestrial biosphere, it seems that plans to build Type I AIs living in terrestrial ecological milieus would be *easier* to achieve via the carbon-based path. The latter leads back to the key idea of living xenobots made out of frog cells [21]. Secondly, in case people still intend to develop an alternative new lifeform based on silicon (which is in theory *not* impossible) that would e.g. inhabit a different potentially shielded artificial ecological niche on Earth, they would still be *at least* confronted with the non-trivial task of implementing the following requirement: MoI understanding – specifically tailored *to that particular shielded artificial ecological niche.*

Present-day Type I intelligent systems do not instantiate any type of MoI *understanding* that would parallel cells in any way. The purpose for which the hardware on which the AI runs has been built, did not aimed itself at realizing a motile co-existence of non-conscious Type I hardware based on chemically modulated excitability. Instead, it was more related to procedures involving mathematical calculations to improve the motile co-existence *of socio-psycho-techno-physical Type II entities.* To put it very simply, in the xenobot studies, the researchers discovered that although the goal specified for the evolutionary algorithms (that served as basis to later biologically realize the xenobots) was *fast locomotion*, collectively *coordinated* contractions emerged *spontaneously* [21] in cells that in turn facilitated locomotion. In general, excitability actions of cells include not only e.g. spatial navigation, photo- and chemotaxis, cell fusion and active feeding by engulfment [127]. Additionally, many eukaryotes, cable bacteria and biofilms engage in *escape* responses and action potentials [127] and many cells can continue with excitable actions in *novel* microfluidic environments [124]. It is clear that the silicon-based hardware utilized in AI projects nowadays does not instantiate knowledge of the described type in any biochemical environment. As long as a non-MoI-like shortcut to CBI is not explained, it appears that current intelligent systems and Type I AIs running on silicon-based hardware lag *qualitatively* behind bacterial biofilms and xenobots when it comes to *understanding.* While these two life forms could be mapped to step 3 of the cyborgnetic ladder (CBI), present-day non-living Type I AIs including intelligent systems are still at step 1.

To recapitulate, instead of dismissing the silicon-based path to life on the grounds of it being "highly unlikely" and akin to science-fiction plots [90] including the Star Trek series, a cyborgnetic assessment must acknowledge that it is *not* impossible. Reasoning about the technical feasibility of things cannot be justified via probabilities. Indeed, justifications are impossible in general and bold conjectures need not be justified at all [56]. As suggested by Deutsch in constructor theory via the possibility-impossibility dichotomy [46], anything that is not prohibited by the laws of nature is possible. Firstly, there is no law of nature prohibiting silicon-based life in the entire universe. Hence, it is possible. Secondly, it could also have been launched *artificially* by an advanced Type II civilization which could have succeeded at it. Thirdly, note also that researchers already implemented an artificial synthesis of molecules comprising *both silicon and carbon* [77]. However, while not impossible in theory, (partially) *silicon*-based artificial Type I life is currently simply *not* there on Earth. Moreover, even if (partially) silicon-based Type II substrates would be in theory possible, it is important to note that those would still have to instantiate *substrate-independent* requisite information from *all* the socio-psycho-techno-physical strata of which Type-II entities are composed of – as encoded in the cyborgnetic ladder of understanding. As described, neither present-day non-living computer hardware nor the software of Type I AI fulfill this requirement. In brief, concerning artificial Type II life *from scratch* – silicon-based or not – it literally is as hard as requisite variety to construct... a new universe.

Chapter 6

Type II *Cynetbit*coin Blockchain

6.1 Motivation

Some of the fictive Cynamian scientists that were members of the simulation committee in Cynam were alarmed by the epistemic fiasco described in Chapter 2. After having analyzed and understood the Type-I-AI-related epistemic intricacies introduced in Chapter 3, discussed in Chapter 4 and deepened in Chapter 5, those scientists began to deliberate over the future of science in Cynam. Firstly, they realized that they *cannot* rely on the inherently epistemically misguided[1] even if well-intended Type-I-AI-based schemes such as deepfake detection, quality prediction, truthful AI, misinformation detection or fact verification. Indeed, as explained in Chapter 4, the epistemic aim of science *cannot* be truth itself nor truer explanations – for lack of direct access to truth from the stance of a "knower". For this reason, it is impossible to train a Type I AI able to detect true or truer *scientific* statements[2]. Would one nevertheless attempt to do so by ill-advisedly declaring that all currently instated scientific theories are true, it would moreover lead to a stagnation of the scientific enterprise in old reputation-based frameworks whose contents could long have been refuted and replaced by better innovative ones. Such a scientific environment would be higly hostile towards statistical outliers and would inherently suppress the diversity of thoughts next to automatically implementing a self-sabotage that could lead to its own death and related existential risks. Secondly, understanding that the epistemic aim of science must instead be the achievement of *better novel* explanatory blockchains (EBs), the fictive Cynamian scientists suddenly comprehended to what extreme extent the *monetary* value of new EBs has been neglected. In Section 6.2, I unravel why.

[1] For a recall, see especially the details provided in Chapter 4.2.2.

[2] Nothing can ever be proved by experiment. It is only in the context of closed fixed mathematical worlds that one can heuristically utilize a notion of true and false (which many may contest even there). When a theory is "proven" mathematically, it only means that given the collated information, it seems to be internally consistent. It can never certify that the theory is true in the real world.

6.2 Type II *Cynetbit*coin

6.2.1 Aim

Crucially, with the epistemic aim of science being the creation of new EBs, I postulate that the latter must entail *the solution of genuine problems* (as e.g. implied in the G_1 operation specified in the exemplary recipe for new EBs displayed in Figure 5.1). From a cyborgnetic perspective, a problem is genuine if it relates to a plausible *harm* case. Hence, one could state that the aim of cyborgnetic science must be to create novel (directly or indirectly) *experimentally falsifiable*[3] EB-based solutions to mitigate harm. (Note that *inherently*, the aim of cyborgnetic philosophy is the decryption or encryption of novel EB-based solutions to mitigate harm while the aim of cyborgnetic art is the encryption of novel EB-based solutions to mitigate harm.) Because new EBs consist of a chain of entangled *Type II cynetbits* [5, 8], one could accordingly state that the currency of science is made of Type II cynetbits. Then, for reasons of epistemic security (see Section 4), the scientific enterprise could be consciously reframed as a marketplace where both the means of payment *and* the product being sold is the "Type II *cynetbit*coin[4]" – which must per definition be in the service of Type II entities by being intrinsically focused on cyborgnetic *harm mitigation.* Because Type II cynetbits cannot be forged (neither by Type I nor by Type II entities), Type II cynetbitcoins *cannot* be generated by an end-to-end-Type-I-AI-pipeline and the scheme is robust against deepfake science attacks.

6.2.2 Scientific Type II Cynetbitcoin Blockchain

Firstly, to avoid attacks on scientific peer review (SPR) availability, one needs *EB-tailored* AI-based plagiarism checks [7] on submissions for candidate new EBs. In case of protest, a Type II entity can possibly analyze the case. Secondly, to mitigate attacks on SPR confidentiality, all submissions must be double-blind. Thirdly, as defense against integrity attacks, one can harness the conjunction of: 1) the explanatory intrusion prevention system (IPS) test [7] followed by 2) a Type-I-falsification-peer-review [7] (Type-I-FPR). Only after passing *both* stages can a candidate new EB be converted to a Type II cynetbitcoin and be added to the *Type II cynetbitcoin blockchain* for science (Type II Sci-CyB). Because all refutations are only provisional, a once instated Type II cynetbitcoin is somehow never deleted – even if refuted and forgotten at a later stage. The most recently appended cynetbitcoin in an area is the currently best available EB in that specific area.

[3] Which means such that can be (provisionally) made problematic by experiment.

[4] One *new* EB consists of multiple cynetbits. However, since those cynetbits are inseparably *entangled*, one would map *one* new EB to *one* Type II cynetbitcoin. In practice, one would be able to buy new EBs of the present with the new EBs of the past – i.e. buying Type II cynetbitcoins of the present using Type II cynetbitcoins of the past. However, not without specific boundary conditions.

Weaker Defense: Explanatory IPS Test

Not interactive; blind text-based setting with normalization to avoid dependency on linguistic style; best available Type I language AI needed to augment the evaluator, one Type II monitor, one Type-I-aided Type II evaluator and one participant needed; positive test corroborates that the text generated by the participant was harder-to-vary than the best currently available Type I AI in that domain (but the test is weak because it does not corroborate that the text was a new EB); negative test means that from the perspective of the evaluator, the text was not even better than the outputs of the currently best available Type I language AI. **Nota bene:** Explanatory IPS test generation should always be monitored by a Type II entity that must be different from the evaluator of that test (otherwise the evaluator would obviously know where to find the submitted text). In this way, one can avoid the generation of too easy samples by the Type I AI and counteract the possibility of adversarial attacks e.g., via a poisoning of datasets where Type II attackers could insert text-level backdoors to let their self-selected papers go through.

Stronger Defense: Explanatory IPS Test + Type-I-FPR

Semi-interactive; blind text-based setting with normalization to avoid dependency on linguistic style; best available Type I language AI needed to augment evaluator A; one Type II monitor, one Type-I-aided Type II evaluator A, one Type II evaluator B for Type-I-FPR part (evaluator A could be the same or alternatively different from evaluator B) and one participant needed; first explanatory IPS test which – only if successful – is then followed by an interactive peer-review where the evaluator must create novel EB-based objections to the text of the participant; positive test corroborates that a new EB was generated (and by extension that participant is of Type II); HOWEVER negative test does NOT corroborate Type-I-ness (i.e. negative test does not say anything about the nature of the participant, only that the evaluator could not identify a new EB – participant could be Type I or Type II). **Nota bene:** Explanatory IPS test generation should always be monitored by a Type II entity that must be different from the evaluator of that test (see explanations specified under the last "Nota bene" point above).

Summary of Basics

After initial security tests, a candidate new EB can only be transformed into a Type II cynetbitcoin and appended to the Type II Sci-CyB *after* a two-staged EB-based selection mechanism combining an explanatory IPS test with a subsequent Type-I-FPR. A candidate new EB is analogous to a scientific paper submission. The Type II Sci-CyB is the collective EB-based scientific knowledge base whose entries are somehow never deleted.

6.2.3 Quantum-Inspired Refinements for Type II Sci-CyB

Recently, researchers proposed a new form for a theoretically feasible but not yet widely implemented *quantum blockchain* [107]. This approach seems to lead to security-relevant advantages in comparison to a classical blockchain due to *time* entanglement (rather than spatial entanglement) phenomena. From the perspective of cynet information theory, this is interesting. Firstly, recall that cynet information theory [8, 5] focuses on three types of superinformation: 1) physical Type I superinformation (mostly referred to as quantum information), 2) socio-psycho-techno-physical Type I superinformation (applicable to *new* non-EB-like linguistic information) and 3) socio-psycho-techno-physical Type *II* superinformation (based on *new* EBs made of Type II cynetbits as described in Section 6.2). The latter must *also* encode a variant of time entanglement – which may represent the basis for the discussed *temporal* cynet shortcut postulated in the context of the explanatory IPS test [8]. Given this commonality between the idea of a quantum blockchain (QB) and new EBs being the constituents of a Type II Sci-CyB, I analyze how one could refine a Type II Sci-CyB with concepts loosely inspired by a QB. This may offer a starting point for a potential series of cross-pollination possibilities that one could deepen in future work.

Quantum Random Number Generator (QRNG)

Within the so-called "verification" protocol of the mentioned QB approach, the authors propose the use of *"a low level sub-algorithm involving a quantum random number generator"* [107] (abbreviated with QRNG in the following). Before elaborating on that, I emphasize that in the Type II Sci-CyB context, we cannot know whether a new EB is true. Instead, when taken together, the two complex stages of Type II Sci-CyB assess whether a candidate submission could correspond to a better EB in comparison to old EBs. Thereby, in the first stage, we are even solely assessing whether the blocks of the submission are better in comparison to those generated by the Type I language AI. However, interestingly, the use of a QRNG e.g. for the assignment of diversified explanatory IPS tests (multiple alternatives for each individual submission) to evaluators but also for Type-I-FPR assignments could become highly beneficial for science – obviously under the condition that all participants agree on a shared rigorous epistemology for new EBs as for instance displayed in Figure 4.1. Among others, it would allow a tunneling through the space of ideas – promoting the diversity of solutions.

Measures Against Tampering

An evaluator that did not yet understand the agreed upon epistemic total order or that maliciously aims at sabotaging the explanatory IPS test could *not* reliably tamper with the process since it may become apparent at a certain point that the evaluations are entirely unrelated to other participants. One also has the possibility to see to what extent this is the case (e.g. whether there were only slight deviations or whether the results seem entirely random). Namely, as in the QB case, the explanatory IPS test offers the possibility to identify the *"ideal target state"* [107] of honest parties – the exact combination of paragraphs forming a sufficiently distinguishable submission. Thus, it may make sense that evaluators who did not retrieve any new EB in the explanatory IPS stage for a comparatively long time could then be selected for the Type-I-FPR round in order to enforce the creation of novel adversarial EBs. For quality managment purposes, the latter could then sometimes be itself encoded as explanatory IPS test (which importantly always includes both anonymization and normalization of linguistic style) presented to other evaluators to identify whether the outputs of the evaluator in question are still sufficiently distinguishable from Type-I-AI-generated alternative streams. Finally, to constrain any *non*-EB-based guesswork, one could *time-lock* the explanatory IPS test in various ways.

Spooky Superinformation Features?

In the mentioned QB design [107], it is possible for photons *that never co-existed* to share entanglement and non-classical measurement correlations. The authors elegantly describe their *"encoding procedure as linking the current records in a block, not to a record* ***of*** *the past, but linking it to the actual record* ***in*** *the past, a record which does not exist anymore*" [107]. Applied to Type II Sci-CyB, I speculate that in analogy, somehow, the new EBs of the present are entangled with all old EBs that ever existed in the past – including those that are long forgotten and literally never co-existed with the new ones. Perhaps, it goes all the way back to what cynet information theory could call the primordial new EB – simultaneously being the most generic one: *the cyborgnetic ladder of understanding* [8] itself. The same could extend to all yet unknown future and beyond that more generally to all counterfactual new EBs. In turn, this could perhaps provide the basis for the impossibility to delete new EBs (see also the cyborgnetic no-deleting theorem [8]). Future work could ideally create new EBs on that subject. It would for instance be intriguing to formulate fundamental theoretical differences concerning the security of a Type-*II*-superinformation-based Type II Sci-CyB in comparison to a *classical* arguably Type-I-AI-imitable peer review process.

QB-based Type II Sci-CyB?

It relates to the idea of *encoding new EBs into time-entangled qubits* [10] for an explanatory IPS test. How can we interpret and even better *practically harness* the parallels with the undecidable black hole information paradox [8]? I may address this in future work.

6.3 Synopsis

In the future, a Type II Sci-CyB model for science could perhaps reward reviewers with Type II cynetbitcoins such that it allows for instance: 1) own publications at correspondingly reduced fees for accepted new EBs, 2) the purchase of past EB-based scientific books and scientific papers that are not open source, 3) the purchase of any other scientifically relevant material from online tutorials to items for experiments over creativity-stimulating software including language AI. On the whole, Bitcoin is grounded in mining by Type *I* substrates solving *Type-I-solvable* new mathematical riddles as "proof-of-work" with *no* direct relation to cyborgnetic harm mitigation. Type *II Cynetbit*coin would be grounded in mining by Type *II* substrates solving *Type-II-only*-understandable cyborgnetic *harm* use cases as "corroboration-of-Type-II-ness" via new EBs. In a Type II Sci-CyB, the product *is* the currency. However, one must not forget that while new EBs solve problems, they also point at new problems. Crucially, one must not forget the dual-use aspect of new EBs. Overall, for instantiated Type II entities, one cannot escape the idea that the price of security is eternal creativity [4]...

In light of the QB-inspired aspects from Section 6.2.3, one can now retrospectively reassess *subtle* security aspects within the two exemplary *strong* peer review strategies that are able to instantiate the so-called Type-II-Cynetbit Protocol [10] (T2CyP) enabling *cyborgnetic signatures* – which are inherently a part of the Type II Sci-CyB framework presented in this chapter. Initially, the Type-I-falsification-event-test [7] (Type-I-FE-test) peer review strategy taken alone was implicitly considered to be as strong as the conjunction of *explanatory IPS test + Type-I-FPR* as peer review strategy. However, it now seems that the latter could be more secure than the Type-I-FE-test because of the explicit usage of *time* entanglement. Thus, conveniently, the scheme of the Type II Sci-CyB depicted in this chapter in Section 6.2.2 seems to already instantiate the recommendable since more secure option. Future work could investigate further in that direction.

Chapter 7

Conclusion

It is imposible for Type I entities to reliably create new EBs[1]. This statement can be made problematic by experiment. However, arguably, since a Type *I deepfake science generator* able to reliably generate new EBs – which I repeat is deemed to be *impossible* on my account – would represent an unprecedented existential risk (see Chapter 4.4), a cyborgnet A succeeding at it may destroy the basis for its own future existence. This is *not* because the Type I AI would be extremely "intelligent". No, it is because malicious *cyborgnets* could harness that Type I universal problem solver at will – also against cyborgnet A then having become a solvable new problem... **Change of scene.** Meanwhile in Cynam, Cyland, the Cynamian Jocker has started to train Type I language AI on old EBs respecting a rigorous epistemic total order. They say he plans a use thereof for a Type *I* artificial general *imitator*[2] to fuel "epistemic babble" [115] fears[3]. They say he also intends to act against the core EB instated in this book just to have fun in case it works... and he has many friends. (To be continued if the laws of nature permit it...)

[1] Those referring to Type I artificial *general* superintelligence must first experimentally falsify the formulated scientific statement and then attempt to (provisionally) refute it via a better new EB. It is not enough to interpret recent AI developments as continuous progress towards superintelligent Type I A*G*I. None of those "progresses" relate to new EBs and everything else can be forged. I suspect that something akin to one or even two law(s) of nature may prohibit Type I entities to reliably create new EBs. I leave this open for future work (but see also [10]). In short, could a Type I AI be super*intelligent*? Perhaps, but intelligence is *not* part of the argumentation utilized here and that Type I AI could *not* be an *artificial general creativity* [10] – for what a cyborgnet (and thus Type-*II*-ness) is required. Thus, a superintelligent Type I AI would *not* be general, it could at most only be a general *imitator* – and the creation of new EBs is *not* imitable.

[2] Humanity must proactively counteract the deployment of any robotic Type I artificial general imitator deployed in conventional real-world environments as the latter could in practice appear indistinguishable from any Type II entity that does not create new EBs in the given setting.

[3] Following cyborgnetic philosophy of science, we neither inhabit a post-truth era nor a post-falsification era [13]. Hence, "epistemic babble" [115], *cannot* be the loss of the ability of people to *tell* the difference between truth and fiction presented as truth. That could never have been our epistemic aim since always impossible. We could and can still strive for better new EBs and distinguish worse from better EBs.

Acknowledgements

This book is dedicated to the potential of cyborgneticity.

I thank Dr. Leon Kester for having reviewed this book and for the contributions as co-author on Chapter 3 and on Chapter 4.

Appendices

Appendix A

Cyborgnetic Metaphysics – A Third Theorem

A.1 Third Impossibility Theorem

Henceforth, the following three impossibility theorems labelled with names stemming from the new cyborgnettish language are instated in cyborgnetic *metaphysics* (which represents a branch of cyborgnetic philosophy):

1. *Maje-theorem:* It is impossible for the laws of nature *not* to be expressible in terms of (encrypted) explanatory blockchains (EBs).
2. *Jauè-theorem:* It is impossible to reliably postdict and predict reliably *hidden* tuples mapping authors to the contents of *novel* EBs they generate.
3. *Mashau-and-Shamela-theorem*[1]*:* It is impossible that self-recreatable self-re-creativity could *not* be omnipresent. However, it is impossible that self-recreatable self-re-creativity *could be* omniscient *in the classical sense*[2].

[1] This bipartite theorem represents a sort of what one could describe as "fraternal-twin-theorem" since it relates to superficially conceptually similar but fundamentally *not* identical notions.

[2] This holds although *all* knowledge that could possibly exist, is embedded in self-recreatable self-re-creativity. The key here is that *to be* is stronger than *to know*. "To know" as described by humans involves a time-like split, a division, a measurement with a *classically* expressible result. To know can diminish.

Bibliography

[1] D. Aerts and L. Beltran. Are words the quanta of human language? Extending the domain of quantum cognition. *Entropy*, 24(1):6, 2021.

[2] D. Alba. Facebook Discovers Fakes That Show Evolution of Disinformation. `https://www.nytimes.com/2019/12/20/business/facebook-ai-generated-profiles.html`, 2019. The New York Times; accessed 04-August-2020.

[3] N.-M. Aliman. Explanatory Blockchain Forgery? `https://www.nadishamarie.jimdo.com/cyborgnetics/`, 2020. Online; accessed 01-September-2021.

[4] N.-M. Aliman. *Hybrid cognitive-affective Strategies for AI safety.* PhD thesis, Utrecht University, 2020.

[5] N.-M. Aliman. *Anagrammatic Singularities – The Cyborgnetic Clockmaker.* Kester, Nadisha-Marie, 2021.

[6] N.-M. Aliman. CA 009: The Non-EB-Like OODA Loops of Type I Intelligent Systems. `https://nadishamarie.jimdo.com/cyborgnetics/`, 2021. Online; accessed 15-September-2022.

[7] N.-M. Aliman. *Cyborgnetics – The Type I vs. Type II Split.* Kester, Nadisha-Marie, 2021.

[8] N.-M. Aliman. *Self-Folding Cynet Worlds – The Ladder and Tali's Paradox.* Kester, Nadisha-Marie, 2021.

[9] N.-M. Aliman. Endnotes UEF. `https://nadishamarie.jimdo.com/uef/`, 2022. Online; accessed 01-June-2022.

[10] N.-M. Aliman. *Self-Climbing Cynet Tree – Hidden Entropy in The Biosphere.* Kester, Nadisha-Marie, 2022.

[11] N.-M. Aliman and L. Kester. Artificial creativity augmentation. In *International Conference on Artificial General Intelligence*, pages 23–33. Springer, 2020.

[12] N.-M. Aliman and L. Kester. Malicious Design in AIVR, Falsehood and Cybersecurity-oriented Immersive Defenses. In *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 130–137. IEEE, 2020.

[13] N. M. Aliman and L. Kester. Epistemic Defenses against Scientific and Empirical Adversarial AI Attacks. In *CEUR Workshop Proceedings*, volume 2916. CEUR WS, 2021.

[14] N.-M. Aliman and L. Kester. *Immoral programming: What can be done if malicious actors use language AI to launch 'deepfake science attacks'?*, pages 179–200. Wageningen Academic Publishers, 01 2022.

[15] N.-M. Aliman and L. Kester. *Moral Programming.* Wageningen Academic Publishers, 2022.

[16] N.-M. Aliman, L. Kester, P. Werkhoven, and R. Yampolskiy. Orthogonality-based disentanglement of responsibilities for ethical intelligent systems. In *International Conference on Artificial General Intelligence*, pages 22–31. Springer, 2019.

[17] N.-M. Aliman, L. Kester, and R. Yampolskiy. Transdisciplinary AI Observatory—Retrospective Analyses and Future-Oriented Contradistinctions. *Philosophies*, 6(1):6, 2021.

[18] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

[19] W. R. Ashby. *An introduction to cybernetics.* Chapman & Hall Ltd., 1957.

[20] G. Aubrun, L. Lami, C. Palazuelos, and M. Plávala. Entanglement and superposition are equivalent concepts in any physical theory. *Physical Review Letters*, 128(16):160402, 2022.

[21] P. Ball. Living robots. *Nature materials*, 19(3):265–265, 2020.

[22] L. Barham and D. Everett. Semiotics and the Origin of Language in the Lower Palaeolithic. *Journal of Archaeological Method and Theory*, 28(2):535–579, 2021.

[23] A. B. Barron and C. Klein. What insects can tell us about the origins of consciousness. *Proceedings of the National Academy of Sciences*, 113(18):4900–4908, 2016.

[24] D. Blackiston, E. Lederer, S. Kriegman, S. Garnier, J. Bongard, and M. Levin. A cellular platform for the development of synthetic living machines. *Science Robotics*, 6(52):eabf1571, 2021.

[25] W. Brooker. *Batman unmasked: Analyzing a cultural icon.* Bloomsbury Publishing USA, 2013.

[26] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.

[27] V. Bufacchi. Truth, lies and tweets: A consensus theory of post-truth. *Philosophy & Social Criticism*, 47(3):347–361, 2021.

[28] G. Cabanac, C. Labbé, and A. Magazinov. Tortured phrases: A dubious writing style emerging in science. Evidence of critical issues affecting established journals. *arXiv preprint arXiv:2107.06751*, 2021.

[29] A. Cadeddu, E. K. Wylie, J. Jurczak, M. Wampler-Doty, and B. A. Grzybowski. Organic chemistry as a language and the implications of chemical linguistics for structural and retrosynthetic analyses. *Angewandte Chemie International Edition*, 53(31):8108–8112, 2014.

[30] Y. Cao, C. Xiao, B. Cyr, Y. Zhou, W. Park, S. Rampazzi, Q. A. Chen, K. Fu, and Z. M. Mao. Adversarial sensor attack on lidar-based perception in autonomous driving. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pages 2267–2281, 2019.

[31] N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 3–14, 2017.

[32] L. Cavalcante Siebert, M. L. Lupetti, E. Aizenberg, N. Beckers, A. Zgonnikov, H. Veluwenkamp, D. Abbink, E. Giaccardi, G.-J. Houben, C. M. Jonker, J. van den Hoven, D. Forster, and R. Lagendijk. Meaningful human control: actionable properties for AI system development. *AI and Ethics*, pages 1–15, 2022.

[33] K.-S. Chae, S.-C. Kim, H.-J. Kwon, and Y. Kim. Human magnetic sense is mediated by a light and magnetic field resonance-dependent mechanism. *Scientific Reports*, 12(1):1–11, 2022.

[34] Y. Chen, X. Yuan, J. Zhang, Y. Zhao, S. Zhang, K. Chen, and X. Wang. Devil's whisper: a general approach for physical adversarial attacks against commercial black-box speech recognition devices. In *Proceedings of the 29th USENIX Conference on Security Symposium*, pages 2667–2684, 2020.

[35] B. Chesney and D. Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 107:1753, 2019.

[36] S. Coghlan and K. Leins. “Living Robots”: Ethical Questions About Xenobots. *The American Journal of Bioethics*, 20(5):W1–W3, 2020.

[37] G. Corera. UK spies will need artificial intelligence - Rusi report. `https://www.bbc.com/news/technology-52415775`, 2020. BBC; accessed 08-November-2020.

[38] M. Cortês, S. A. Kauffman, A. R. Liddle, and L. Smolin. Biocosmology: Biology from a cosmological perspective. *arXiv preprint arXiv:2204.09379*, 2022.

[39] M. Cortês, S. A. Kauffman, A. R. Liddle, and L. Smolin. Biocosmology: Towards the birth of a new science. *arXiv preprint arXiv:2204.09378*, 2022.

[40] M. Cortês, S. A. Kauffman, A. R. Liddle, and L. Smolin. Biocosmology: Towards the birth of a new science. *arXiv preprint arXiv:2204.09378*, 2022.

[41] L. Da Costa, P. Lanillos, N. Sajid, K. Friston, and S. Khan. How active inference could help revolutionise robotics. *Entropy*, 24(3):361, 2022.

[42] J. Degrave, F. Felici, J. Buchli, M. Neunert, B. Tracey, F. Carpanese, T. Ewalds, R. Hafner, A. Abdolmaleki, D. de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, 2022.

[43] J. Deng, Z. Yang, D. Samaras, and F. Wang. Artificial Intelligence in Drug Discovery: Applications and Techniques. *arXiv preprint arXiv:2106.05386*, 2021.

[44] M. J. Dennis, G. Ishmaev, S. Umbrello, and J. Van den Hoven. Values for a Post-Pandemic Future. In *Values for a Post-Pandemic Future*, pages 1–19. Springer, 2022.

[45] D. Deutsch. *The beginning of infinity: Explanations that transform the world.* Penguin UK, 2011.

[46] D. Deutsch. Constructor theory. *Synthese*, 190(18):4331–4359, 2013.

[47] D. Deutsch. The logic of experimental tests, particularly of Everettian quantum theory. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 55:24–33, 2016.

[48] D. Deutsch and C. Marletto. Constructor theory of information. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2174):20140540, 2015.

[49] C. Doss, J. Monschein, D. Shu, T. Wolfson, D. Kopecky, V. A. Fitton-Kane, L. Bush, and C. Tucker. Deepfakes and Scientific Knowledge Dissemination. *Europe PMC*, 2022.

[50] M. R. Ebrahimkhani and M. Levin. Synthetic living machines: A new window on life. *Iscience*, page 102505, 2021.

[51] O. Evans, O. Cotton-Barratt, L. Finnveden, A. Bales, A. Balwit, P. Wills, L. Righetti, and W. Saunders. Truthful AI: Developing and governing AI that does not lie. *arXiv preprint arXiv:2110.06674*, 2021.

[52] C. Fields, K. Friston, J. F. Glazebrook, and M. Levin. A free energy principle for generic quantum systems. *Progress in Biophysics and Molecular Biology*, 2022.

[53] H.-C. Flemming, J. Wingender, U. Szewzyk, P. Steinberg, S. A. Rice, and S. Kjelleberg. Biofilms: an emergent form of bacterial life. *Nature Reviews Microbiology*, 14(9):563–575, 2016.

[54] L. Floridi. Artificial intelligence, deepfakes and a future of ectypes. *Philosophy & Technology*, 31(3):317–321, 2018.

[55] L. E. Foley, R. J. Gegear, and S. M. Reppert. Human cryptochrome exhibits light-dependent magnetosensitivity. *Nature communications*, 2(1):1–3, 2011.

[56] D. Frederick. *Against the Philosophical Tide: Essays in Popperian Critical Rationalism.* Critias Publishing, 2020.

[57] D. Frederick. Critique of Brian Earp's writing tips for philosophers. *Think*, 20(58):81–87, 2021.

[58] D. Frederick et al. The Contrast Between Dogmatic and Critical Arguments. *Organon F*, 22(1):9–20, 2015.

[59] D. Frederick et al. Falsificationism and the Pragmatic Problem of Induction. *Organon F*, 27(4):494–503, 2020.

[60] M. D. Fricker, L. L. Heaton, N. S. Jones, and L. Boddy. The mycelium as a network. *The fungal kingdom*, pages 335–367, 2017.

[61] J. Fridman, L. F. Barrett, J. B. Wormwood, and K. S. Quigley. Applying the theory of constructed emotion to police decision making. *Frontiers in psychology*, 10:1946, 2019.

[62] A. P. Gieseke. " The New Weapon of Choice": Law's Current Inability to Properly Address Deepfake Pornography. *Vanderbilt Law Review*, 73(5):1479–1515, 2020.

[63] GPT-2 (pre-trained). Text Generation API. `https://deepai.org/machine-learning-model/text-generator`, 2021. Online; accessed 31-March-2021.

[64] X. Han, Y. Hu, L. Foschini, L. Chinitz, L. Jankelson, and R. Ranganath. Deep learning models for electrocardiograms are susceptible to adversarial attack. *Nature medicine*, 26(3):360–363, 2020.

[65] K. Hao. The Biggest Threat of Deepfakes Isn't the Deepfakes Themselves. `https://www.technologyreview.com/2019/10/10/132667/the-biggest-threat-of-deepfakes-isnt-the-deepfakes-themselves/`, 2019. MIT Technology Review; accessed 08-November-2020.

[66] K. Hartmann and K. Giles. The Next Generation of Cyber-Enabled Information Warfare. In *2020 12th International Conference on Cyber Conflict (CyCon)*, volume 1300, pages 233–250. IEEE, 2020.

[67] K. Hartmann and C. Steup. Hacking the AI-the Next Generation of Hijacked Systems. In *2020 12th International Conference on Cyber Conflict (CyCon)*, volume 1300, pages 327–349. IEEE, 2020.

[68] C. J. Hernández-Castro, Z. Liu, A. Serban, I. Tsingenopoulos, and W. Joosen. Adversarial machine learning. In *Security and Artificial Intelligence*, pages 287–312. Springer, 2022.

[69] S. Hussain, P. Neekhara, M. Jere, F. Koushanfar, and J. McAuley. Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3347–3356. IEEE, 2021.

[70] S. Jain, V. M. Cachoux, G. H. Narayana, S. de Beco, J. D'alessandro, V. Cellerin, T. Chen, M. L. Heuzé, P. Marcq, R.-M. Mège, et al. The role of single-cell mechanical behaviour and polarity in driving collective cell migration. *Nature physics*, 16(7):802–809, 2020.

[71] G. Jakubowski. What's not to like? Social media as information operations force multiplier. *Joint Force Quarterly*, 3:8–17, 2019.

[72] J. Jiménez-Luna, F. Grisoni, N. Weskamp, and G. Schneider. Artificial intelligence in drug discovery: Recent advances and future perspectives. *Expert Opinion on Drug Discovery*, pages 1–11, 2021.

[73] A. Johanson. Modern topology and Peirce's theory of the continuum. *Transactions of the Charles S. Peirce Society*, 37(1):1–12, 2001.

[74] A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. Tygar. *Adversarial machine learning*. Cambridge University Press, 2018.

[75] F. Jotterand and C. Bosco. Keeping the "human in the loop" in the age of artificial intelligence. *Science and Engineering Ethics*, 26(5):2455–2460, 2020.

[76] N. Kaloudi and J. Li. The AI-based Cyber Threat Landscape: A Survey. *ACM Computing Surveys (CSUR)*, 53(1):1–34, 2020.

[77] S. J. Kan, R. D. Lewis, K. Chen, and F. H. Arnold. Directed evolution of cytochrome c for carbon–silicon bond formation: Bringing silicon to life. *Science*, 354(6315):1048–1051, 2016.

[78] B. Katherine. Envisioning Our Posthuman Future: Art, Technology and Cyborgs, 2015.

[79] S. Kauffman. Is There a 4th Law for Non-Ergodic Systems That Do Work To Construct Their Expanding Phase Space? *arXiv preprint arXiv:2205.09762*, 2022.

[80] S. Kauffman and A. Roli. The world is not a theorem. *Entropy*, 23(11):1467, 2021.

[81] S. A. Kauffman and A. Roli. The Third Transition in Science: Beyond Newton and Quantum Mechanics–A Statistical Mechanics of Emergence. *arXiv preprint arXiv:2106.15271*, 2021.

[82] F. C. Keil. Explanation and understanding. *Annu. Rev. Psychol.*, 57:227–254, 2006.

[83] Y. Kim, F. Bertagna, E. M. D'Souza, D. J. Heyes, L. O. Johannissen, E. T. Nery, A. Pantelias, A. Sanchez-Pedreño Jimenez, L. Slocombe, M. G. Spencer, et al. Quantum biology: An update and perspective. *Quantum Reports*, 3(1):80–126, 2021.

[84] D. Kirat, J. Jang, and M. Stoecklin. Deeplocker–concealing targeted attacks with AI locksmithing. *Blackhat USA*, 1:1–29, 2018.

[85] S. Kriegman, D. Blackiston, M. Levin, and J. Bongard. A scalable pipeline for designing reconfigurable organisms. *Proceedings of the National Academy of Sciences*, 117(4):1853–1859, 2020.

[86] S. Kriegman, D. Blackiston, M. Levin, and J. Bongard. Kinematic self-replication in reconfigurable organisms. *Proceedings of the National Academy of Sciences*, 118(49):e2112672118, 2021.

[87] H. Kwon. Dual-Targeted Textfooler Attack on Text Classification Systems. *IEEE Access*, 2021.

[88] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[89] B. Leberecht, D. Kobylkov, T. Karwinkel, S. Döge, L. Burnus, S. Y. Wong, S. Apte, K. Haase, I. Musielak, R. Chetverikova, et al. Broadband 75–85 MHz radiofrequency

fields disrupt magnetic compass orientation in night-migratory songbirds consistent with a flavin-based radical pair magnetoreceptor. *Journal of Comparative Physiology A*, 208(1):97–106, 2022.

[90] D. Lincoln. Misconceptions of Science: Is Silicon-based Life Possible? `https://www.thegreatcoursesdaily.com/misconceptions-of-science-is-silicon-based-life-possible/`, 2021. FROM THE LECTURE SERIES: Understanding the Misconceptions of Science; accessed 23-September-2021.

[91] X. Liu, K. Wan, Y. Ding, X. Zhang, and Q. Zhu. Weighted-sampling audio adversarial example attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4908–4915, 2020.

[92] J. Lu, H. Sibai, and E. Fabry. Adversarial examples that fool detectors. *arXiv preprint arXiv:1712.02494*, 2017.

[93] T. Mahlangu, S. January, T. Mashiane, M. Dlamini, S. Ngobeni, and N. Ruxwana. Data poisoning: Achilles heel of cyber threat intelligence systems. In *Proceedings of the ICCWS 2019 14th International Conference on Cyber Warfare and Security: ICCWS*, 2019.

[94] J. Mallatt, M. R. Blatt, A. Draguhn, D. G. Robinson, and L. Taiz. Debunking a myth: plant consciousness. *Protoplasma*, 258(3):459–476, 2021.

[95] V. Mannalath, S. Mishra, and A. Pathak. A comprehensive review of quantum random number generators: Concepts, classification and the origin of randomness. *arXiv preprint arXiv:2203.00261*, 2022.

[96] D. Noble. The illusions of the modern synthesis. *Biosemiotics*, pages 1–20, 2021.

[97] R. Noble and D. Noble. Was the watchmaker blind? Or was she one-eyed? *Biology*, 6(4):47, 2017.

[98] P. Oltermann. ' At first I thought, this is crazy': the real-life plan to use novels to predict the next war. `https://www.theguardian.com/lifeandstyle/2021/jun/26/project-cassandra-plan-to-use-novels-to-predict-next-war`, 2021. The Guardian; accessed 15-September-2022.

[99] H. Öztürk, A. Özgür, P. Schwaller, T. Laino, and E. Ozkirimli. Exploring chemical space using natural language processing methodologies for drug discovery. *Drug discovery today*, 25(4):689–705, 2020.

[100] N. R. Pace. The universal nature of biochemistry. *Proceedings of the National Academy of Sciences*, 98(3):805–808, 2001.

[101] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman. Sok: Security and privacy in machine learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 399–414. IEEE, 2018.

[102] D. K. Picariello. *Politics in Gotham: The Batman Universe and Political Thought*. Springer, 2019.

[103] K. Popper. *The poverty of historicism*. Routledge, 2013.

[104] K. Popper. *Conjectures and refutations: The growth of scientific knowledge*. Routledge, 2014.

[105] K. Popper and W. W. Bartley III. *Realism and the aim of science: From the postscript to the logic of scientific discovery*. Routledge, 2013.

[106] A. Prindle, J. Liu, M. Asally, J. Garcia-Ojalvo, and G. M. Süel. Ion channels enable electrical communication in bacterial communities. *Nature*, 527(7576):59–63, 2015.

[107] D. Rajan and M. Visser. Quantum blockchain using entanglement in time. *Quantum Reports*, 1(1):3–11, 2019.

[108] J. Ramsay and D. R. Kattnig. Radical triads, not pairs, may explain effects of hypomagnetic fields on neurogenesis. *arXiv preprint arXiv:2206.08192*, 2022.

[109] P. Ranade, A. Piplai, S. Mittal, A. Joshi, and T. Finin. Generating fake cyber threat intelligence using transformer-based models. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2021.

[110] L. Reynolds and K. McDonell. Multiversal views on language models. *arXiv preprint arXiv:2102.06391*, 2021.

[111] J. Rodriguez, T. Hay, D. Gros, Z. Shamsi, and R. Srinivasan. Cross-domain detection of gpt-2-generated technical text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1213–1233, 2022.

[112] J. Rohrlich. Romance Scammer Used Deepfakes to Impersonate a Navy Admiral and Bilk Widow Out of Nearly $300,000. `https://www.thedailybeast.com/romance-scammer-used-deepfakes-to-impersonate-a-navy-admiral-and-bilk-widow-out-of-nearly-dollar300000`, 2020. Daily Beastl; accessed 08-November-2020.

[113] R. Sarimov, V. Binhi, and V. Milyaev. The influence of geomagnetic field compensation on human cognitive processes. *Biophysics*, 53(5):433–441, 2008.

[114] N. Schick. *Deep fakes and the infocalypse: What you urgently need to know*. Hachette UK, 2020.

[115] E. Seger. The greatest security threat for the post-truth age . `https://www.bbc.com/future/article/20210209-the-greatest-security-threat-of-the-post-truth-age`, 2021. BBC; accessed 14-September-2022.

[116] E. Seger, S. Avin, G. Pearson, M. Briers, S. Ó. Heigeartaigh, H. Bacon, H. Ajder, C. Alderson, F. Anderson, J. Baddeley, et al. Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world. *The Alan Turing Institute*, 2020.

[117] J. M. Shainline. Does cosmological evolution select for technology? *New Journal of Physics*, 22(7):073064, 2020.

[118] L. Slocombe, M. Sacchi, and J. Al-Khalili. An open quantum systems approach to proton tunnelling in DNA. *Communications Physics*, 5(1):1–9, 2022.

[119] L. D. Smith, F. T. Chowdhury, I. Peasgood, N. Dawkins, and D. R. Kattnig. Driven spin dynamics enhances cryptochrome magnetoreception: Towards live quantum sensing. *arXiv preprint arXiv:2206.07355*, 2022.

[120] C. Stupp. Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. `https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402`, 2019. The Wall Street Journal; accessed 04-August-2020.

[121] J. Takhar, H. R. Houston, and N. Dholakia. Live very long and prosper? Transhumanist visions and ambitions in 2021 and beyond..., 2022.

[122] T. Thellefsen and B. Sorensen. *Charles Sanders Peirce in his own words: 100 years of semiotics, communication and cognition*, volume 14. Walter de Gruyter GmbH & Co KG, 2014.

[123] P. Tully and L. Foster. Repurposing Neural Networks to Generate Synthetic Media for Information Operations. `https://www.blackhat.com/us-20/briefings/schedule/`, 2020. Session at blackhat USA 2020; accessed 08-August-2020.

[124] L. Tweedy, P. A. Thomason, P. I. Paschke, K. Martin, L. M. Machesky, M. Zagnoni, and R. H. Insall. Seeing around corners: Cells solve mazes and respond at a distance using attractant breakdown. *Science*, 369(6507), 2020.

[125] F. Urbina, F. Lentzos, C. Invernizzi, and S. Ekins. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3):189–191, 2022.

[126] W. P. Walters and R. Barzilay. Critical assessment of AI in drug discovery. *Expert Opinion on Drug Discovery*, pages 1–11, 2021.

[127] K. Y. Wan and G. Jékely. Origins of eukaryotic excitability. *Philosophical Transactions of the Royal Society B*, 376(1820):20190758, 2021.

[128] L. Wang, L. Zhou, W. Yang, and R. Yu. Deepfakes: A new threat to image fabrication in scientific publications? *Patterns*, 3(5):100509, 2022.

[129] M. Wang and R. A. Dean. Movement of small RNAs in and between plants and fungi. *Molecular plant pathology*, 21(4):589–601, 2020.

[130] P. Werkhoven, L. Kester, and M. Neerincx. Telling autonomous systems what to do. In *Proceedings of the 36th European Conference on Cognitive Ergonomics*, pages 1–8, 2018.

[131] G. Woo. Downward counterfactual search for extreme events. *Frontiers in Earth Science*, 7:340, 2019.

[132] J. Xu, L. E. Jarocha, T. Zollitsch, M. Konowalczyk, K. B. Henbest, S. Richert, M. J. Golesworthy, J. Schmidt, V. Déjean, D. J. Sowood, et al. Magnetic sensitivity of cryptochrome 4 from a migratory songbird. *Nature*, 594(7864):535–540, 2021.

[133] K. Xu, G. Zhang, S. Liu, Q. Fan, M. Sun, H. Chen, P.-Y. Chen, Y. Wang, and X. Lin. Adversarial T-shirt! Evading person detectors in a physical world. In *European conference on computer vision*, pages 665–681. Springer, 2020.

[134] R. V. Yampolskiy. Preface: Introduction to AI Safety and Security. *Artificial Intelligence Safety and Security*, 2018.

[135] X. Yuan, P. He, Q. Zhu, and X. Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9):2805–2824, 2019.

[136] H. Zadeh-Haghighi and C. Simon. Magnetic field effects in biology from the perspective of the radical pair mechanism. *arXiv preprint arXiv:2204.09147*, 2022.

[137] F. M. Zanzotto. Human-in-the-loop Artificial Intelligence. *Journal of Artificial Intelligence Research*, 64:243–252, 2019.

[138] T. Zhang. Deepfake generation and detection, a survey. *Multimedia Tools and Applications*, 81(5):6259–6276, 2022.

[139] B. Zhao, S. Zhang, C. Xu, Y. Sun, and C. Deng. Deep fake geography? When geospatial data encounter Artificial Intelligence. *Cartography and Geographic Information Science*, pages 1–15, 2021.

[140] A. Zhavoronkov, Y. A. Ivanenkov, A. Aliper, M. S. Veselov, V. A. Aladinskiy, A. V. Aladinskaya, V. A. Terentiev, D. A. Polykovskiy, M. D. Kuznetsov, A. Asadulaev,

et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature biotechnology*, 37(9):1038–1040, 2019.

[141] Z. Zhou, D. Tang, X. Wang, W. Han, X. Liu, and K. Zhang. Invisible mask: Practical attacks on face recognition with infrared. *arXiv preprint arXiv:1803.04683*, 2018.

9 789083 226576