Generic Analyses for AI, Safety and Security Research

Cyborgnetics – The Type I vs. Type II Split

Dr. Nadisha-Marie Aliman, M.Sc.

Cyborgnetics –
The Type I vs. Type II Split

Cyborgnetica –

De Type I vs. Type II Splitsing

GENERIEKE ANALYSES VOOR KI, VEILIGHEID EN BEVEILIGING

Utrecht, Netherlands

Dedicated to †Dipl. Ing. Ramani Aliman, †Pearl Kuruneru and †Vasantha Kuruneru

Contents

Chapter 1

Introduction

Cyborgnetics is a new generic meta-discipline whose aim is to systematically facilitate the documentation, critical analysis and mitigation of any socio-psycho-techno-physical *harm* studied in (applied) science, engineering and/or philosophy. It is applicable to a wide variety of safety and security domains ranging from cybersecurity over criminology to safety in virtual reality (VR). Besides, it may also be for instance transferable to theory formation in psychology and cognitive science pertaining to harm constructions. However, this book focuses on applications of cyborgnetics to contemporary problems in AI research, in AI safety and security (including links to VR), as well as in cybersecurity engineering. In a nutshell, cyborgnetics applies *cyborgnet theory* to conjectured harm events, analyses those rigorously and develops countermeasures – which includes to be introduced *cyborgnetic creativity augmentation* techniques. Though only very briefly mentioned in this book, cyborgnetics is embedded in an epistemological bedrock denoted *unbound(ed) epistemic funambulism*. The latter comprises an own philosophy (of science) and an own metaphysical framework complemented by an artistic stance which conceives of art as a procedure of encryption – and *not* expression – performed by a generic fictional entity.

Pertinent safety and security issues such as "deepfakes", AI attacks and adversarial AI including implications for science and engineering are not only analyzed through a new lens but also used as stepping stone for novel conjectures. Among others, I reflect upon new alternatives to classical Turing Tests to shield against either Type I AI itself or against "non-explanatory" potentially Type-I-AI-generated *contents*. I provide tentative answers and formulate novel questions pertaining to the difference between Type I and Type II entities, but also specifically between human and non-human *great apes*. Whilst this book is a written monologue conducted for purposes of self-education as an end in itself, it could perhaps as a side effect stimulate the creativity of a few technology and ethics practitioners or any interested party maneuvering the increasingly complex harm landscape. Perhaps, it could provide creativity-augmenting inputs for those asking what could be unique about *us* amidst advances of present-day AI potentially fuelling doubts.

Cyborgnet theory (introduced in Chapter 3 and abbreviated with CT) transforms the old substrate-dependent question on what is unique about humans into a *substrate-independent* quest of explaining and understanding the features of what *we* have in common. While "we" can seem to be just a word, so is the potentially illusory infinity of differences that we could exhibit depending on the perspectives we take. Under CT, "we" is a *generic template*, an infinite *potential* of what one could term cyborgneticity (i.e. a potential of cyborgnets and cyborgnet networks). To put it very simply, a *cyborgnet* corresponds to a directed graph of active nodes with the two following properties: 1) the graph comprises at least one Type II entity and 2) there exists at least one Type II entity in that graph which has altered function (e.g. via restoration, enhancement, physical impact, affective change or even deterioration) due to the additional integration of at least one Type I entity (e.g. of artificial, technological, ideational, procedural, biological nature). Cyborgnets are the principal unit of CT analyses utilized for the critical development of *harm narratives* whose nature is not only descriptive but crucially also *explanatory*. In cyborgnetics, language is seen as primary technology interwoven in all socio-psycho-techno-physical strata that could be affected by studiable harm. As can be extracted from Chapter 2 and Chapter 4 to 10, to *understand* and to specifically harness *language AI* and *language* itself may be key to tackle harm in the deepfake era – be it in the context of AI safety and security, cybersecurity or co-creation in social virtual reality.

In complex multicausal harm domains, CT analyses examine *structured* intra-cyborgnetic and inter-cyborgnetic harm narratives which can represent individual composite scenarios or nested and extended sequences thereof. In addition, both retrospective and future-oriented *counterfactual* deliberations, i.e. a consideration of worse scenarios that *could have happened but did not* and better ones that *could but did not yet happen* are an inherent part of the analytical CT method. Procedurally, a CT analysis first taxonomically models recent or open problems in a retrospective descriptive analysis (RDA). In a second step called retrospective counterfactual risk analysis (RCRA), plausible downward counterfactual scenarios projecting to the immediate past are crafted based on the mentioned RDA problems. In a third and final step denoted future-oriented counterfactual defense analysis (FCDA), a CT analysis then formulates explanation-anchored practical solutions projecting to plausible upward counterfactuals of the near future which are proposed to tackle both RDA and RCRA problem clusters. In the main, while the meta-discipline of cyborgnetics – the generically formulated application of CT – could be classified as a scientific or engineering-related endeavor of harm mitigation whose FCDAs are among others strongly focusing on feedback-loops fostering security awareness and cyborgnetic creativity augmentation, it may also be indirectly but intimately linked to ethics. Firstly, whilst the price of security is eternal creativity [10], security implies to be free from danger and threat – i.e. harm. Secondly, following the theory of dyadic morality, it holds that morality is intrinsically pluralistic but based on a universal *harm*-based cognitive template reflecting *"[...] cognitive unity in the variety of perceived harm"* [232].

In the following, I summarize the main contributions of this book:

1. In Chapter 2, I ask whether it is impossible in the long-term to shield from misguiding Type-I-AI-generated text inputs in social media spaces. My answer is no. I introduce a novel theoretical method for an interactive asymmetric text-based *Type-I-shield* which is formally neither equivalent to bot detection nor to any Turing Test.

2. In Chapter 3, I present a new framework for the systematic study of socio-psycho-techno-physical *harm* – I call it *cyborgnet theory* (CT). It is a *substrate-independent* view whose stratified focal points are the so-called *cyborgnets* composed of Type I and Type II active nodes embedded in *structured* network dynamics. I explain why I depart from past ontologies of technology-related harm and show how CT improves upon those (in particular stratigraphy of harm but also actor-network-theory).

3. In Chapter 4, 5 and 8, I elaborate on *content-centered* defense methods against novel technically feasible but not yet prevalent *text-based* forms of what one could term *deepfake science* attacks. Chapter 4 also proposes epistemic defenses against deepfake text in cybersecurity (especially related to *deepfake cyber threat intelligence*).

4. In Chapter 6, I reassess the old anthropological question on the nature of the difference between human and non-human *great apes*. Using CT, I provide new perspectives on how *creativity* evolved from Type I great apes to universal cyborgnets of Type II. I try to answer the question on whether any *Type II AI* project is known and whether humans are the only Type II entities on this planet.

5. In Chapter 7, I utilize epistemic stratagems from Chapter 5 to develop a complementary *double deception technique* that could be employed in cybersecurity to deter intellectual property theft performed by cyberattackers (be it in cyberespionage or ransomware contexts). I explain how it improves upon a past alternative solution.

6. In Chapter 9, I reflect upon a new Type-I-AI-related harm use case that I denote *honey mind trap* (HMT). I carry out a reappraisal of the defense strategy of a Type-I-shield introduced in Chapter 2. I explain why in practice, i.a. due to the *cyborgnetic dilemma*, new *cyborgnetic creativity augmentation* measures that could also be transferred to co-creation in *social virtual reality* represent a better alternative to defend against HMT attacks in the near future. In Chapter 10, I briefly entertain philosophical thought experiments explaining why *we* could *not* be HMTs.

7. In Chapter 11, I provide an overview on the novel (currently five) *impossibility theorems of cyborgnetics* that were implicitly or explicitly stated in this book.

Outline

- Chapter 2 introduces the *Type-I-falsification-event-test* to corroborrate (which is not equivalent to proving) the Type-II-ness of a test subject in a one-by-one setting.
- Chapter 3 presents cyborgnet theory, a generic analytical framework of *epistemic*, *cybernetic* and *cybersecurity-oriented* nature conceived to support practical procedures of documenting, examining and counteracting *harm* in complex multi-causal problem domains.
- Chapter 4 elucidates *generic* epistemic defenses against scientific and empirical adversarial AI attacks (abbreviated with *SEA AI attacks*) and illustrates their instantiation via two use cases: scientific writing and cyber threat intelligence.
- Chapter 5 describes the novel idea of an *explanatory intrusion prevention system* preceeding peer-review to shield against contents that are not hard-to-vary enough in comparison to the best available state-of-the-art Type I language AI.
- Chapter 6 elaborates on why complex relational transformations pertaining to the artificial augmentation of *creativity* in a counterfactual symbolic landscape contributed to fundamental differences in information processing between the different *species* of non-human and human great apes. Though, I explain why there may be very few isolated exceptions to this pattern that already occured in *individuals*.
- Chapter 7 introduces EXPLANATORY-FORGERY, a document- and paragraph-level double deception technique to deter intellectual property theft committed by cyberattackers.
- Chapter 8 explains why it is impossible to reliably craft counterfeits of the so-called *explanatory blockchains* (irrespective of the nature of the agent).
- Chapter 9 discusses Type-I-AI-related mind perception biases as vulnerabilities against *honey mind traps* (HMTs) and explains how to defend against those.
- Chapter 10 contains philosophical reflections on HMTs in the context of the new epistemic bedrock of *unbound(ed) epistemic funambulism*.
- Chapter 11 concludes, provides an overview of the current five *impossibility theorems of cyborgnetics* and compiles compact take-home messages for the deepfake era.
- Chapter 12 discusses ideas for future research.

Chapter 2

Self-Shielding Worlds

This chapter with a title encoding at least a double meaning has been written for purely autodidactic purposes as by-product to another project and as fragmented temporary mental clipboard. It is based on a slightly modified form of the essay that I uploaded to the website `https://nadishamarie.jimdo.com/clipboard` on November 23, 2020.

2.1 The Practical Problem: Social Bots

In recent years, the topic of social media bots with automated accounts designed to emulate certain human behavioral patterns [42] emerged as an issue of international relevance. Social bots can be maliciously instrumentalized for AI-enhanced disinformation operations [114, 132]. For instance, they can be used to qualitatively manipulate contents [275], to quantitatively influence the discourse landscape (e.g. by artificial likes and shares influencing trending topics [132] and steering collective attention), to bind human users in homophilic peer groups [205] via affective contagion effects [89, 132] or to spread AI-generated fake material [14]. (Beyond that, automated social media accounts for espionage [223] disguised with AI-generated fake profile pictures have been generated whereby synthetic AI-generated face images are perceived as more authentic [256] due to their intrinsic characteristic as being designed to mimic average features. As a consequence, humans can exhibit higher social conformity [256] towards these fake persona increasing the potential of manipulation or the uptake of maliciously tailored contents.) Generally, it has been stated that social bots represent a risk to *"[...] public opinion, democracy, public health, stock market and other disciplines"* [188].

In the long-term, next to the use of social bots to automate large-scale disinformation which could even cause civil wars [230], it is also thinkable that social bots could be harnessed to automate social engineering, sextortion and also *"[...] harassing an at-*

risk teenager into suicide, ruining personal relationships, or inducing a victim into doing something financially or personally risky" [42]. In addition, a thereby widely understudied risk represents the future misuse of sophisticated bots to potentially automate the production of fake science articles (see a simple prototype by Yampolskiy conceived in other research contexts [136]) that could e.g. simply confirm (randomly or carefully) selected theories. This risk could be fueled by AI-generated fake experiments or manipulated historical samples (see demonstration of MIT for educational purposes [181]). In the light of pre-existing "fake science" [127] circumstances that academia faces and amidst epistemic threats [88] and post-truth narratives [111] that many empiricists (wrongly [14]) seem to face, fake science bots could add to the so-called *automated disconcertion* [15] pattern. The latter refers to the societal-level epistemic confusion that arises by the mere existence of AI-generated fakery.

In short, maliciously designed bots with natural language capabilities represent a serious threat to social media AI safety and beyond. Hence, bot detection appears a highly valuable defense endeavor. Ideally, a "bot shield" could be implemented that would allow for bot-free spaces in social media where explicitly desired by users. However, many researchers agree that: 1) AI-aided bot detection is destined to fail in the long-term and 2) even humans may in the long-term lose their ability to distinguish automated bots from humans due to the ability of the former to become better and better at the imitation game. In short, many assume that bot detection must fail in the long run since bots will ultimately pass a Turing test. (Thereby, the imitation game underlying e.g. generative adversarial networks is viewed as a sort of weak version of the Turing Test [42].) While bot detection may indeed not represent a long-term solution, I briefly explain in this chapter that the notion of a one-by-one shield facilitating the exclusion of systems like automated bots must not necessarily fail. I provide an exemplary theoretical solution in which a *human* tester performs a discerning test – that is however *not* analogous to an imitation game and comes with certain novel caveats. Thereby, the epistemology of Deutsch [73] is relevant.

2.2 A Theoretical Solution

2.2.1 Type I vs. Type II Systems

As proposed in [10], I distinguish between two disjunct types of systems: Type I and Type II systems. Type II systems are all systems for which it is possible to *consciously create and understand explanatory knowledge*. Type I systems are all systems for which this is an impossible task. All Type II systems are conscious. A small subset of Type I systems can be conscious too (think e.g. of non-human mammals). Obviously, *all present-day*

AIs are of Type I and non-conscious. Type II AI is clearly non-existent nowadays (but it is physically in theory possible). The only currently known Type II systems are humans. Automated bots with natural language capabilities are trivially Type I AIs. In the following, I describe how one can utilize this Type I property to conceive of a pragmatic for current purposes *human-executed* test for a *Type-I-shield* (described in Subsection 2.3.1) – a test which can *among others* shield from automated bots. Importantly, this test *cannot* be able to separate Type I from Type II systems. It solely answers the following question: *"did the human tester experience a Type-I-falsification-event in the test subject?"* I thus call this test a Type-I-falsification-event-test (abbreviated with Type-I-FE-test in the following). Importantly, next to a *suitable question*, one requires a *suitable task* and a *suitable domain* of interest *to the test subject* (were it a Type II system).

2.2.2 Type-I-Falsification-Event-Test (Type-I-FE-Test)

Next to an *adversarial* human tester allowed to use any technical aids, I assume a suitable language interface that can be used by human tester and test subject. The test subject can be a bot but also a human. Here, for simplicity, I exemplarily focus on *textual* communication given that current bots are often operating in that modality, but speech-based or even braille-based versions are naturally equally conceivable. In the pre-test phase, the test subject can *self-select a domain of interest* in which he wishes to be tested from a steadily augmented set of real-world domains (if no suitable option is available, the test subject may deposit a request). The Type-I-FE-test task consists of two obligatory subtasks: 1) creating at least one *novel yet unsolved* real-world problem(s) in the chosen domain and presenting it in text form, 2) generating a *hard-to-vary explanation* [73] on how to solve at least one of the self-generated problem(s) and presenting it in text form. The human tester initially provides *a narrow problem cluster* under the selected domain to ease generation. In case the adversarial human tester estimates that the test subject convincingly succeeded in both 1) *and* 2) within a self-determined reasonable amount of time[1], the Type-I-FE-test is marked as positive. The Type-I-FE-test is negative otherwise. The human tester is allowed to (in text form) criticize the material generated and can ask to refine it. Further, the human tester is allowed to singularly repeat a first failed procedure in another domain in concert with the test subject.

2.2.3 Theoretical Implications

- ***Positive Test:*** In the light of the descriptions provided under Subsection 2.2.1, a positive Type-I-FE-test *corroborates* that the test subject is a Type II system. This

[1] In theory, a Type II test subject must be able to reliably repeat such an event if motivated to do so in a self-chosen domain and given enough time.

should not be confused with a proof or a confirmation. It is a corroboration and nothing more. Interestingly, it does also *not* logically entail that the test subject necessarily was a human. It only corroborates that it is a Type II system – *any* Type II system. Were it a Type II AI, one would not be able to tell it apart based on this test. It is only because Type II AI is assumed to be non-existent today, that many people could be inclined to assign a human nature to the test subject. For this reason, the Type-I-FE-test is *substrate-independent* with regard to positive results.

- ***Negative Test:*** A given system can fail at a Type-I-FE-test for multiple reasons – of which only one possible option is that the system is a Type I system. Of course, a negative Type-I-FE-test can mean it is a Type I system. However, it could also mean that it is a Type II system that was unwilling to participate. Or a Type II system whose preferred domain was not yet part of the available domains. Or a Type II system for which it needs more time to be able to elicit a Type-I-FE, simply because it is yet too young. Finally, a negative Type-I-FE-test can also not attest that the system is a Type I *AI*. It only documents that the system did *not* elicit a Type-I-FE: namely, it failed to show a practical sample of how it can consciously create and understand explanatory knowledge in that specific test session in that domain to that specific human tester. The Type-I-FE-test is also *substrate-independent* with regard to negative results. How and why a Type-I-FE-test can still be of high practical value is explained under Subsection 2.3.

- ***Relation to Imitation Game and Turing Test:*** Deutsch already explained [73] why a Turing Test asking *whether a machine can think* (for clarity, I strictly use "think" in this chapter to refer to consciously creating and understanding explanatory knowledge [2]) cannot be equated to an imitation game of behavioral nature. In short, I endorse this view. I regard an imitation game with respect to any conceivable task *except the task mentioned in the Type II system definition* (the task to consciously create and understand explanatory knowledge), as *not* fundamentally inaccessible for a Type I AI. However, concerning this exceptional task of consciously creating and understanding explanatory knowledge (ccuEK in the following), it is impossible for a Type I AI and any other Type I system – irrespective of any level of intelligence. In short, in theory it is permissible that a Type I AI succeds in the imitation game at any conceivable *non-ccuEK* task. But it is ccuEK that is targeted in a Type-I-FE-test and it is clear that for a positive result, an imitation of anyone does not help. Creating novel yet unsolved problems and then providing a hard-to-vary explanation on how one could solve those cannot be learned by a

[2] On purely theoretical grounds, it holds generally that to implement an artificial system able to perform this task that humans are able to emulate must be *possible* in the light of the universality of computation [73].

Type I system. There is no training data for a such a task. There is not even a ground-truth. It is an open-ended task for open-ended updatable domains.

- ***Relation to Turing Question:*** The theoretical question of Turing on *whether a machine can think* will be called Turing Question in the following. (Recall, that for clarity in this specific chapter, I assume that the verb "think" strictly corresponds exactly to the ccuEK task.) I use the term Turing Question to disentangle it from the test devised to assess this question which has been termed the imitation game or the Turing Test. In short, it is not mandatory that the Turing *Question* needs to be necessarily addressed by a test called "imitation game" or "Turing Test". Coming to the Type-I-FE-test, it is clear that it *cannot* answer the Turing Question – already because this test is substrate-independent while *the Turing Question is substrate-dependent*. A positive Type-I-FE-test can corroborate that the test subject thinks, but does not provide any information on whether it is a human or an AI or a cyborg or a Type II alien. Generally, Deutsch explains why for a human to know whether a specific machine thinks, someone would actually need *to explain this human how this machine has been built* and only once the human then *understood* it, will he consider the question to be answered [73]. Imagine the case in which one is told that a person one knows since five years is a Type II AI. While it could seem at first sight that a sort of individualized Turing Question has now been answered, it is not the case since one could not shake the idea that it simply is a human. In order to really grasp it, one would need high-level explanatory insights into the way it has been developed. But as Deutsch explains [73], would someone have provided these explanations *before* one would have ever seen this system, one would already know the answer to the question – making any subsequent test called Turing Test or imitation game obsolete. In short, it is harder for someone to know whether you are *a machine that thinks* than to know that *you* think – even if you are a machine. Overall, the Turing Question seems in the end to be reducible to a testless Turing *Explanation* on *how to build Type II AI* – which is not known yet.

2.3 Practical Use of Theoretical Solution

2.3.1 Type-I-Shield

As mentioned earlier, a negative Type-I-FE-test does *not necessarily* signify that the test subject is a bot and not a Type II system. It could also mean for instance that it is a Type II system unwilling to participate. However, for the practical purposes of e.g. near-term bot communication checks in social media given the negative impacts illustrated in Section 2.1, it might be suitable since in order to use the service, human users would

be *mostly willing to pass the test* – especially *if they can independently select a preferred domain of interest*. Note also that thinkable test domains are not limited to science and can also include arts, morality, philosophy, literature and so on. For purely practical interactive reasons, it is thus reasonable to assume that adult humans will be not only willing but also able to bring about a positive Type-I-FE-test *corroborating their Type II nature*. By achieving that in practice, the Type-I-FE-test can help to create a pragmatic *Type-I-shield* of limited information content but high social use. The main point here is that in theory, due to the view on ccuEK, *no single automated bot would be able to achieve a positive Type-I-FE-test*. Against this background and given the motivation of human users to create bot-free spaces, the fact that a negative test can have multiple interpretations is not that weighty anymore. In short, a Type I shield assigns test subjects to two possible groups: a first group with a positive test consisting *solely* of systems whose Type II nature is thereby corroborated (but not proven) and a second group of systems with a negative test that can comprise a hybrid ensemble of e.g. Type I AI, slightly too young Type II systems, Type II systems whose special interests have yet to be integrated into the available domain options for the test, cats and dogs (being biological Type I systems), unwilling Type II systems and so forth. To recapitulate, *a social media space that would solely comprise entities with a positive Type-I-FE-test would be shielded against Type I systems* – while still risking to having *not yet* included some remaining Type II systems. The latter problem can i.a. be addressed by letting these Type II systems formulate requests for novel domains to be integrated in the steadily growing pool of domain options. For too young human test subjects, the problem would dissolve with time. For remaining cases, human creativity could further improve upon that since there is *no fundamental* obstacle.

2.3.2 Test Engineering Method for Type-I-Shield

To experimentally realize the abstract test setup described in Subsection 2.2.2, I briefly describe how one could craft a simple fit-for-purpose setting that could be harnessed in social media AI safety. Recently, proactive research in AI safety [17, 128] and in security topics at the intersection of AI and virtual reality (AIVR) [15] proposed to utilize co-creation design fictions (DFs) to develop strategies. In particular, in order to craft tailored *defense methods* against malicious design in AIVR safety, it has been suggested to craft them on the basis of DFs grounded in *threat models* [42] known from cybersecurity. For illustration purposes, I briefly comment on how one could then process a Type-I-FE-test taking AIVR safety as an exemplary *domain* and disinformation via VR deepfakes [15] as *narrow problem cluster* under that domain. The first Type-I-FE-test subtask of creating at least one novel yet unsolved real-world problem could be a written DF narrative for a plausible threat model. The second subtask of providing a hard-to-vary explanation on how to solve this problem would be a short written proposal on a corresponding defense

method and hard-to-vary explanation on why it is suited against that specific threat model. How exactly the test subject generates these elements mentally is left open. Some may want to mentally conceive of it as a future projection, others would frame it as a downward counterfactual of the past or as free search. Since the human tester is allowed to criticize the generated contents, one must only be ready to clarify one's statements.

2.3.3 Practical Caveats and Side Effects

The database of domains needs to be steadily updated and augmented. Each domain would need to be continuously renewed by a growing and refined population of narrow problem clusters associated to it. For instance, in diverse domains of scientific nature one could imagine the scope of a narrow problem cluster to be situated within the scope of an average research article. One important requirement is that *each particular single instance of a narrow problem cluster* is understood as a *single use* element excluded from further consideration to be able to meet the condition of the first subtask in the Type-I-FE-test (namely to create a novel yet unsolved real-world problem). Proactively, techniques similar to plagiarism assessment tools (here to avoid duplicate entries and fulfill the first condition of the test) could be already used additionally for those critical cases in which a narrow problem cluster needs to be re-used. While a multiplicity of instances can be associated to such a cluster, one may be more at risk to generate an already existing instance. In addition, future work needs to address how to train the human testers on critical thinking to avoid epistemic mistakes grounded in flawed mind perception of Type I AI. The two Type-I-FE-test subtasks need to be obligatory and automatable negotiation attempts to evade e.g. the second subtask should be seen through. Beyond that, questions of intellectual property might need to be thoroughly addressed. Many might view each Type-I-FE-test as potentially encoding intellectual property while some participants may prefer to opt for anonymity. Ideally, it could give rise to a criticism-oriented feedback-loop e.g. if the solutions provided at the end of each positive Type-I-FE-test are further criticized and novel narrow problem clusters are then consequently added to the database. On the whole, in scientific domains such a modus operandi could be a form of participatory science, in arts collaborative artwork, in philosophy a novel collective discourse form. Indeed, a side effect of the described Type-I-shield application in practice could be to augment human creativity [13].

2.4 Conclusion

In this very brief chapter utilized as ephemeral mental clipboard representing a by-product to another project, I discussed a possible theoretical solution for a simple one-by-one so-

cial media Type-I-shield based on a human-executed Type-I-FE-test – that is crucially *not* analogous to any imitation game or Turing test. The Type-I-FE-test simply asks *whether a human tester was able to experience a Type-I-falsification-event in the test subject*. Thereby, the idea is that such an event would corroborate the latter's ability to consciously create and understand explanatory knowledge, a feature which can then be used for a Type-I-shield. In order to bring about such an event in practical settings, I proposed a twofold test setup where the test subject after having selected a domain of preference, is first faced with the subtask of generating a new problem in that domain and second with the further subtask of providing a hard-to-vary solution to that self-generated problem. As exemplary *concrete* test engineering method to realize this abstract test setup, I mentioned the design fiction use case for the generation of proactive defense methods in AI(VR) safety [15]. As opposed to classical bot detection tools, a positive Type-I-FE-test *corroborates* the substrate-independent Type II nature of the test subject while a negative test result is also substrate-independent and can importantly have multiple causes by what it is *not* equivalent to bot detection. While given a negative test result the test subject could have been a bot, it could also e.g. alternatively be linked to the unwillingness of a human test subject to participate.

To alleviate the latter possibility in practice, I argued that the very fact that many people may be interested in establishing bot-free spaces (given the negative socio-psycho-technological impacts that maliciously designed social bots already cause as described in Section 2.1), could lead to a higher willingness to bring about positive Type-I-FE-test results which could become a shared motif. Overall, the Type-I-FE-test assigns test subjects to two separate groups: a first homogeneous group composed *solely* of systems for which their Type II nature has been corroborated (i.e. a *Type-I-free* group) and a second potentially heterogeneous group of systems that did not bring about a Type-I-FE-event in that specific test session in that domain with that human tester. The higher the motivation of social media users to create bot-free spaces, the less Type II systems could be contained in the latter heterogeneous group making the Type-I-shield a profitable measure for more and more users. Such types of tests could be used on multiple reasonable occasions for safety-aware online social gatherings. In short, a future practical framework for Type-I-FE-tests could create *self-shielding worlds* of increasing cognitive diversity [214] crafted *by* Type II systems *for* Type II systems – but much work needs to be done for theory to meet practice.

2.5 Future Work

Future work could reflect about the possibility to develop a *critical test* building on certain ideas mentioned in this chapter. For instance, the Type-I-FE test discussed may only be

meaningful in case a *fundamental* difference of ability between Type I and Type II systems (related to the ccuEK task) actually holds. For instance, from the perspective of certain AI researchers, this difference may seem not to represent a matter of kind but rather a matter of degree – especially for those who believe that the ccuEK task could be mastered by imitation. For those entities, a possibility to make the presented paradigm problematic in an experimental setting, would be to implement an AI that would be able to – *without any conscious understanding of explanations* – repeatedly bring about a positive Type-I-FE-test. (Candidates could be e.g. (future extensions of) GPT-3 [47].) In this case, the paradigm would still not be terminally falsified, since one could by way of example still argue ad hoc that the Type-I-FE test does not actually capture the essence of the ccuEK task (e.g. is inferior to this task) or that the particular considered implementation of the test was not valid. Alternatively, one could argue that a hypothesized inability to master the ccuEK task is principally not falsifiable experimentally which would exclude it from scientific scrutiny and reframe it instead as a predominantly philosophical question. Further, among the external factors to check, one could try to identify whether a human adversary inadmissibly interfered with the test setting to fool the human tester. In any case, a hard-to-vary explanation on why and a description on how it happened will be needed. Nevertheless, it seems in principle conceivable.

However, if the assumption is that the Type-I-FE-test represents a suitable proxy for the ccuEK task and no explanations for external disturbing factors can be found under repeated experimental efforts (while at the same time better explanations are available on how the AI has been programmed to be able to deceive convincingly), one may at a certain point need to reject the following: 1) the paradigm suggesting a categorical difference between Type I and Type II systems, 2) the practical utility of a Type-I-FE-test and 3) the idea that virtual social spaces are "shieldable" from Type I sophisticated bots. It would then be difficult to escape what one could call the *imitation paradigm* – a framework postulating that all human-performed tasks including the ccuEK task can be mastered by a Type I AI by imitation[3]. While the imitation paradigm is obviously

[3]For the sake of clarity, note that from a purely theoretical perspective, it is still possible that (in the case the Type-I-FE-test is considered to represent a suitable proxy to the ccuEK task in the subtle and peculiar way described in this chapter) another *as yet unknown* method *differing* from the imitation paradigm could exist that could reliably and repeatedly bring about positive Type-I-FE-tests for a Type I AI i.e. *without any conscious understanding of explanations*. Among others, it could for instance perhaps be possible to consistently bring about *illusions* of Type-I-FE events. To give a speculative example, consider extreme applications of future advanced (neuro-)cognitive hacking strategies which could – when enacted by an advanced Type I AI previously maliciously crafted by a human attacker – reliably fool human testers into constructing the judgment or false memory of having experienced a Type-I-FE event. However, in this case, a clarifying explanatory insight on how this strategy was implemented by the human attacker would in itself represent a hard-to-vary explanation on why the paradigm postulating a categorical difference between Type I and Type II systems is *not* falsified by the positive Type-I-FE-test results *under these particular conditions*. Ideally, one could then proceed in creating a novel improved proxy test.

considered to be crucially *wrong* in this chapter, it is helpful to facilitate a critical stance towards prior assumptions and formulate those in a way that they *could* be made problematic experimentally. However, once the imitation paradigm would be embraced, one would also need to assume that explanation-based and criticism-centered science can be automatized. Consequently, one would *from then on* face a specific form of epistemic threat [88] realized generally in the routine of scientific peer-review but also specifically applied to this chapter: *"was it written by a Type I AI?"* But *"modern humans are distinctive among animals for using tools as symbols"* [26] and I propound that humans are also distinctive among animals for using *symbols as tools* – apt for a doubly ambiguous *artificial creativity augmentation* [13].

2.6 Contextualization

This chapter provided an introduction into a long series of deliberations on how to shield against epistemic distortions achieved via textual Type-I-AI-generated inputs. However, a certain number of concepts may require a further elucidation. For instance, the notion of explanatory knowledge utilized by Deutsch [73] still appears too vague while the idea of the presented Type-I-shield is *entity-centered* and may intrinsically lead to unintentional exclusions. For this reason, I perform further refinements. Firstly, Chapter 5 introduces novel more precise ontological distinctions pertaining to what could be meant by explanatory knowledge. Secondly, Chapter 9 provides a reappraisal and explains why an *open* and *content-centered* approach is superior to the entity-centered Type-I-shield. So far, the book only provided an implicit glimpse on how the modus operandi of *cyborgnetics* could look like. Cyborgnetics is a scientific and engineering-related endeavor involving the systematic application of cyborgnet theory (CT) to corresponding problems. CT can provide assistance in documenting, examining and *counteracting* complex safety and security issues across a wide range of contexts. Generally, CT can serve as a supportive *generic* frame of reference applicable to a multiplicity of disciplines tackling different varieties of *socio-psycho-techno-physical harm*. For instance, CT could be applied to areas ranging from criminology to AI safety over security in brain-computer-interface research and extended reality. Methodologically, CT critically analyzes relational, counterfactual and temporally extended *narratives* emerging from *structured* network *dynamics* whose *substrate-independent* and *stratified* focal points are the so-called *cyborgnets* composed of Type I and Type II *active nodes*. In the next Chapter 3, I compactly introduce the concept of a cyborgnet and outline the skeleton of a possible fit-for-purpose modus operandi for CT-based analyses. Thereby, I relate CT to earlier ontologies of technology-related harm from the criminology and zemiology domain explaining the added value of CT from an *epistemic* perspective. CT may reveal profound socio-anthropological implications.

Chapter 3

Cyborgnet Theory

This chapter written for purposes of self-education is based on a slightly modified form of the paper that I uploaded to the website `https://nadishamarie.jimdo.com/clipboard` on March 31, 2021.

3.1 The Practical Problem: Understanding Social Harm in the Context of Technology

In the last decades, technological advancement led to an increasing saliency of the technical aspects implicated in social harms. Against this backdrop, different fields within criminology and zemiology conceptualized multiple novel socio-technical approaches to study technology-related harm and especially crime [259, 273]. In the area of cybersecurity, the harm capacities of malicious actors equipped with sophisticated technological tools already reached international dimensions [68, 157] and defense strategies require a socio-technical contextualization. Similarly, in the field of AI safety [17, 119] and AI ethics [105, 250], different socio-technological paradigms have been suggested to address the challenges of risks in the context of AI systems. Another example are the efforts in the nascent fields of ethics, safety and security for extended reality (XR) where researchers attempt to formulate novel socio-technological strategies to address the complex heterogeneous harms that could emerge from virtual reality, augmented reality and mixed reality applications [12, 55, 70, 191]. Analogous security work started in the field of brain-computer-interface (BCI) research revealing apparently unprecedented intricate socio-technical risks [35, 41]. An additional complication is that risks in different areas addressing technology-related harm can often be exacerbated by combination with each other which impedes the formulation of adequate proactive practices, countermeasures, defense mechanisms and investigation procedures. At another level, already the mere

documentation process of occurring risk instantiations poses serious problems e.g. due to the difficulty to disentangle unintended from intended first-order effects or the challenge to identify the scope of unforeseen second-order effects linked to the opacity and unpredictability of certain technological elements.

As technology becomes more and more complex and its impacts expand in scope, speed and scale, it is important to expose the ontological assumptions underlying frameworks that analyze technology-related harm to critical scrutiny. Generally, it is necessary in order to adjust and update *practical* techniques to acquire a better grip on this convoluted dynamic harm landscape with issues ranging from ethical implications over crimes to existential risks for humanity. In this vein, this chapter introduces *cyborgnet theory* (CT), a novel explanatory ontological framework pertaining i.a. to what is often understood as technology-related harm. Note that while the last paragraph provided a few exemplary domains to which CT can be applied, CT is of *generic* nature and there are numerous old and new domains for which it could be utilized as interpretative lens. A distinct feature to be elaborated further is that CT does not only account for what is classically associated with technology-related harm but vastly extends beyond such cases. While CT can be applied to e.g. cybercriminology, biosafety and biosecurity, security for internet of things (iOT) devices and cyberphysical systems – all of which are modern examples where the technological elements are highly salient, CT can also help to analyze harm in seemingly *"technology-unrelated"* cases studied in e.g. criminology, forensic science, psychiatry, history and anthropology.

In the next Section 3.2, I explicitly introduce different ontological distinctions that are central in the CT framework and form the basis for CT analyses. For illustrative purposes, I then exemplify sense-making under CT by contrasting it with a state-of-the-art approach to technology-related harm from the domain of criminology and zemiology that has been termed "stratigraphy of harm" [273] – which itself extended beyond the well known actor-network theory framework [217]. I discuss commonalities and differences and explain the added value of CT. Thereafter, Section 3.3 briefly elaborates on the generic practical use of CT by specifying aims and limitations. I explain how CT can support transdisciplinary efforts that aim to improve the retrospective documentation and analysis of harm instantiations and develop corresponding proactive practices, countermeasures and defenses across diverse complex multi-causal problem domains. Furthermore, I describe why CT is not only a philosophical meta-theory, but also an albeit simple *scientific* framework whose practical utility could be experimentally tested since its core assumptions are apt to experimental falsification. Finally, in Section 3.4, I conclude and shortly elucidate why CT could reveal far-reaching socio-anthropological implications providing incentives for future work.

3.2 A Theoretical Solution: Cyborgnet Theory

In CT, the main unit of description and explanation is what I refer to as *cyborgnet*. Note that this concept is *not* to be confused with the term "cyborg". A cyborgnet does not only significantly differ from the latter but it also is the case that the term "cyborgnet" is decisively used in a much more *general* and crucially *substrate-independent* way by which it refers to a much broader set of entities. For clarity, before providing an exact definition of a cyborgnet, I first introduce the basic elements it is composed of, namely the so-called *Type I* and *Type II* entities that build its constituting *active nodes*. In Subsection 3.2.1, the ontological distinction between Type I and Type II entities is described. Based on this, Subsection 3.2.2 introduces the notion of active nodes which while sharing some resemblances with actor-network-theory [217] is fundamentally different in relevant epistemic aspects. In Subsection 3.2.3, a definition of the term cyborgnet is assembled by making use of the introduced terminology. Building on this definition, Subsection 3.2.4 elaborates on sense-making under CT by contrasting it with stratigraphy of harm [273] (SoH). Overall, when analyzing harm through the interpretative lens of CT, it is postulated that all social harm implicates *socio-psycho-techno-physical* strata in a way to be described and that common dichotomies (including the human vs. technology divide) can be abandoned, while the substrate-independent dichotomy of Type I vs. Type II entities is epistemically *indispensable*. While SoH attempted a middle ground avoiding social determinism and technological determinism by conceptualizing *technology-related* harm in terms of imbricated strata, CT introduces a *socio-psycho-techno-physical indeterminism* of stratified cyborgnets. In this way, as opposed to SoH, CT is applicable to any type of social harm – even harm that is not classically associated with technology – including retrospective contemplations back to the advent of human abilities to invent complex tools.

3.2.1 Type I vs. Type II Entities

As recently undertaken in AI safety[10], I analogously distinguish between two disjunct types of entities: Type I and Type II entities. Originally, this ontological distinction from the AI safety domain was formulated at the level of systems and thus it classified a given system as either being a Type I or Type II system. The subtle additional element that CT adds to that approach is that the *entities* considered in CT are by no means limited to what one might commonly understand by "system" since an entity can be any "thing" in an inflationary sense. Despite this very general formulation, CT allows a sharp categorization explained in the following. Under CT, Type II entities are all entities for which it is possible to *consciously create and understand explanatory knowledge*. Type I entities are all entities for which this is an impossible task. All Type II entities are

conscious. A small subset of Type I entities can be conscious too (think e.g. of non-human mammals).

Crucially, Type II entities as well as conscious Type I entities in CT need to be instantiated in a physical substrate given that consciousness is linked to the cybernetic control of a physical entity (a process creating a virtual simulacrum of the actively sampled world experienced from an egocentric perspective governed by inherently affective embodied dynamics). Ideas themselves are *non-conscious Type I entities*[1] as is the quasi-totality of artefacts that people currently refer to as "technology" – from systems in virtual reality to tools in biotechnology. Obviously, by extension, *all present-day AIs are of Type I and non-conscious* which also includes automated bots with natural language capabilities but also all sorts of hypothetical non-conscious AI systems that are often imagined when researchers thematize "artificial superintelligence". Type II AI is clearly *non-existent* nowadays. However, from a theoretical perspective its implementation is physically possible in the light of the universality of computation [73].

The only Type II entities currently known to humanity are humans. Note that for those people that consider cyborgs to belong to a distinct category than humans, it makes sense to extend the previous sentence, since the first officially recognized cyborgs which also self-identify with that term already exist today. However, a very important distinctive feature of CT consists in the fact that the coarse mashed ontological distinction underlying the concept of a cyborgnet is substrate-independent i.e. it is irrelevant whether a considered Type II entity is a human, a cyborg, a hypothetical alien or a hypothetical future AI. Further, CT implies that the difference between Type I and Type II entities is fundamental and not a matter of degree. Either an entity has the *ability* to consciously create and understand explanations in which case it is a Type II entity or it has not – making it a Type I entity.

3.2.2 Active Nodes

A cyborgnet is composed of so-called *active nodes*. Active nodes are either Type I or Type II entities. Further constraints on the complex composition of cyborgnets and networks of cyborgnets are introduced in the next Subsection 3.2.3. For now, the focus is first on expounding the meaning of active nodes, the low-level basic units of a cyborgnet. Active nodes form a directed graph indicating unidirectional or bidirectional relations between those nodes. The attribute "active" is used since the existence of a node is conditioned on the presence of *at least* one (unidirectional or bidirectional) *relation* between two entities

[1]Interestingly, this also implies that those ideas which Type II entities construct about other Type II entities are of Type I such that not only imaginary friends are Type I entities, but also the idea of an existing real friend of Type II.

which in any case needs to be physically or/and mentally *enacted* in some way. For this reason, active nodes do never exist in isolation. Beyond that, each instance of an active node is labelled with the category of the entity it represents leading to the distinction between Type I active nodes and Type II active nodes. Note that while similarly to actor-network theory [217] (abbreviated with ANT in the following), CT considers entities in a network as starting point, there are many relevant epistemic differences. First, as opposed to ANT, symmetry is not necessary at the level of active nodes in CT since unidirectional relations between these nodes are permissible. In fact, ANT focuses on interactions, while CT only presupposes *relations in the mathematical sense* [58] of a directed edge which is not necessarily bidirectional. Second, although ANT is stated to consider symmetric interactions, there is an inherent human vs. technology division underlying its ontological distinctions while CT simply distinguishes between Type I vs. Type II active nodes. Third, while ANT assings the general concept of an "actant" to all nodes it considers which has been stated to resemble *"the cyborg concept that unites both human and nonhuman elements"* [259], CT does *not* engage in such assignments since in CT it holds that: 1) there is no human vs. nonhuman distinction in the first place, 2) the emergence of cyborgnets is hierarchically situated *at a layer above* the layer of active nodes. Fourth, the latter reveals a further crucial difference: while ANT analyzes *flat* network topologies, CT is inherently apt to capture structure, depth and *hierarchies* at multiple levels as becomes apparent in the next subsections.

3.2.3 Cyborgnets

A cyborgnet refers to the construct of a directed graph of active nodes with the following requirements: 1) the graph comprises *at least one* Type II entity and 2) *there exists* at least one Type II entity in that graph which has altered function (e.g. restoration, enhancement, affective change or even deterioration) due to the additional integration of *at least one* Type I entity (e.g. of artificial, synthetic, technological, ideational, procedural nature). In CT, cyborgnets are the main unit of analysis utilized for the critical development of descriptive and *explanatory*[2] harm narratives. It can be an unforeseen or goal-oriented situated conceptualization, a local or global construction that can be of ephemeral or more persistent nature across different spatio-temporal and hierarchical scales. Typically, in a CT analysis one would consider an extended sequence of intra-cyborgnet and inter-cyborgnet relations. (For more details on the concrete modus operandi underlying CT analyses, see Subsection 3.3.1.) In the following, the notion of cyborgnet is further clarified and illustrated via processes from human phylogeny and ontogeny exemplifying relevant interplays between active Type I and active Type II entities.

[2]This specific attribute is a highly relevant difference between CT and ANT since the latter is mainly conceived as *descriptive* tool.

To familiarize oneself with the cyborgnet concept, it may be helpful to envisage a minimalistic cyborgnet. For this purpose, I briefly depict the first cyborgnets in human phylogeny. By definition, one requires at least the combination of a Type II and a Type I entity. Since a Type II entity is defined as an entity able to consciously create and understand explanatory knowledge, it is clear that a candidate Type II active node must at least possess *language* abilities without which explanations cannot be formulated. In linguistics, one can identify two main historical accounts of language: a punctuated and a gradualist one. The former view hold by Chomsky [61] assumes a recent abrupt emergence of a unique universal human language faculty characterized by a hierarchical recursive grammar and brought into existence via a genetic mutation in the brain of *homo sapiens* around 50,000 and 70,000 years ago. The latter view is reflected in different types of models all of which consider biocultural co-evolution as gradually developing enabler of language. One notable example is the approach of Barham and Everett [26] which understand language as the communication via symbols i.e. whose *minimum* requirement solely consists of a conjunction of symbols and linear order forming a simple "G1 grammar" [83]. On this view inspired by the semiotics approach of Charles Sanders Peirce [195], language is possible *without* any recursive structures and its origins can be traced back to *homo erectus* as early as 1,000,000 years ago. Decisively for CT, irrespective of which of those accounts holds, one can assume that language originated in a context where at least *stone tool use* was already ubiquitous.

Recall that the notion of Type I entities can encompass any non-Type-II "thing" from ideas over systems to tools and processes. Hence, in the light of the aforesaid, it becomes clear that since the earliest Type II entity in human phylogeny inherently exhibits linguistic abilities *and* inhabits a world permeated by tool use, it instead needs to be understood as the conjunction of *at least* 1) this Type II entity, 2) the Type I entity of language and 3) the Type I entity of material tools. Strictly speaking, given that both language and material tool use are quintessentially *social* activities, the earliest Type II entities were of course *also* surrounded by 4) other human Type II entities. This seems very natural especially from the gradualist standpoint but is also permissible under the punctuated view if the there supposed genetic mutation for the language faculty affected multiple individuals of a population simultaneously. However, note that even in the from my perspective implausible but still conceivable case that a first Type II entity existed as unique specimen amidst Type I biological "conspecifics" because the genetic mutation (in the punctuated view) first only affected *one* individual, the existence of this individual would still have been representable as a conjunction of *at least* 1), 2) and 3). I call this conjunction the *minimal conjunction* of active nodes in CT.

Coming back to the definition of what constitutes a cyborgnet, it becomes apparent that this minimal conjunction from human phylogeny already fulfills the requirements for a cyborgnet. Vygotsky [263] stated that *"through others we become ourselves"* [216]. This

insight is often applied to the human social milieu in which relations between humans form the basis for the development of the self [40]. CT allows a broader perspective on that stament: the "other" need not be a human nor does it need to be a Type II entity. More precisely, Type I entities are other "others" for Type II entities. In short, Type II entities never exist in isolation and can always be analyzed through the lens of a cyborgnet. Particularly, the existence of a Type II entity involves the presence of a split vis-à-vis "Type I others". Under CT thus, any human-technology dichotomy is illusory and ill-phrased – not only because of the substrate-dependent bias. In fact, already the first Type II entities (whether one assumes them to have emerged 50,000 or 1,000,000 years ago) were embedded in cyborgnets *ab initio*. On that view, from an ontological perspective, modern technology since the industrial revolution did *not* make human existence more "cyborgnet-like" than before. Ontologically speaking, a modern cyborg like Neils Harbisson equipped with an eyeborg implanted in his skull [140] exists within a cyborgnet as did the first Type II entity from within its minimal conjunction of active nodes in the Stone Age. The differences lie e.g. in the number of active nodes in cyborgnets, the complexity of the strata they are constituted of, their capacity in terms of speed, memory, scale and scope and in the increasing awareness about this complex stratified hybridity of human existence.

When analyzing cyborgnet origins in human phylogeny, one may notice the interplay between language and stone tools acting as cohesive forces. In the gradualist view, the intimate interaction between language and stone tool-making can be understood via a twofold interpretative lens. On the one hand, diversified and more complex stone tools for foraging and hunting engendered the need for language to enable the tool-making related teaching of *"increasingly complex coordinated actions"* [26] for which communicative gestures became insufficient. On the other hand, language in turn integrated these early material technological artefacts into all aspects of social reality. Stone tools represented a highly salient socio-material affordance and have been conjectured to have served as first symbols given the vast array of associations and connotations their usage may have carried. The usage of symbols and linear sequencing then allowed language as a habitual and productive communication channel extending beyond icons and indexes to which the econiches of other animals are restricted. As stated by Barham and Everett: *"modern humans are distinctive among animals for using tools as symbols"* [26]. CT is in line with this idea and additionally assumes that modern humans are distinctive among animals for using *symbols as tools*. In short, under CT, language itself is i.a. understood as active tool and thus as (non-conscious) Type I entity with among many others vitally also *technological* properties. The next paragraph further illuminates this conjecture and elucidates why CT assumes by deduction that all social harm is i.a. composed of socio-psycho-*techno*-physical dimensions.

An early definition of technology from 1994 suggested by Naughton [183] described it as

"the application of scientific and other knowledge to practical tasks by organisations that involve people and machines". Following CT, a first tentative and much more general substrate-independent definition of technology could be formulated as follows: *the application of explanatory knowledge to practical tasks by one or more cyborgnets*. Even in the punctuated view of language origins, it is assumed that language was from the beginning on used in the service of survival-relevant practical tasks such as the teaching of sophisticated tool-making for hunting e.g. of advanced spears. Such more complex tools required explanatory knowledge including the capacity to invent via systemizing mechanisms [215] which qualitatively differ from the mere associative learning abilities exhibited by certain non-human primates known to utilize simple tools. In sum, language is endowed with technological properties as its historical origins lie in the transmission of sensory-motor and affective simulations from one human Type II entity to another in the context of quintessentially *practical* tasks[3] such as teaching, the coordination of collective activities, efficient allostatic co-regulation and information exchange with conspecifics for which it is still – not exclusively but – habitually used nowadays.

Finally, it is noteworthy that also in human ontogeny, a Type II individual never exists outside the context of a cyborgnet. At the pre-natal stage, the substrate of a future Type II entity is co-embodied [65] i.e. literally physiologically embedded *within* the womb located in the substrate of another Type II entity co-regulating the necessary developmental paths starting in homeostatic regimes. Post-natally, this entity not only inhabits but can now directly interact with a rich socio-psycho-techno-physical environment of cyborgnet affordances – required for it to grow in the light of the strong human allostatic dependency [24] on conspecifics. The growth of this inherently social brain then leads to the

[3]Note that while certain theories imply that the function of language is a form of communication with *other* humans i.e. involving a human sender and a receiver that are different from each other, this seems a view centered on neurotypical humans in habitual situational contexts. For instance, in certain contexts, (auto-)echolalia as used by autistic persons indicates that a practical task for language can also lie outside of this schema. In some cases, language can e.g. fulfill the different practical function of facilitating introspection, learning and/or self-regulation [69] with sender and receiver being two entities but being located within the same individual. As mentioned earlier in this subsection, Type I entities are other "others" to Type II entities. In fact, the self-regulatory effects of *self-talking* (where people engage in monologues with a projected "I" or "You") and covert inner speech have been analyzed in multiple studies. Thereby, in studies performed with neurotypical individuals, positive self-talking (in the "You" form) was able to improve performance in problem solving tasks [79] including self-counselling [237] while inner speech can represent a support for learning, reasoning and creativity. Interestingly, when considering human ontogeny, it has been postulated by Vygotsky [264] that inner speech arises via an internalization of language-based social exchanges (as practiced between children and caregivers) yielding an internalized conversation with the self for the regulation of one's own behavior [8]. At the same time, dysfunctional processes in inner speech generation have been implicated in schizophrenia [7]. Perhaps it is the case in certain schizophrenic experiences that the mental "Type I other" in a specific *local* train of thought is interpreted as being a Type II entity while this seems to simultaneously locally reduce the experiencer to a Type I entity (i.e. somehow an inversion of the split) as seen in "thought insertion" delusions [117].

emergence of a novel additional cyborgnet with i.a. the new Type II entity as active node. In this context, it is important to stress that given its definition, the term cyborgnet can apply to constellations with only *one* Type II entity alongside Type I active nodes but also to coalesce elements from multiple smaller cyborgnets i.e. comprising *multiple* Type II active nodes alongside Type I active nodes. Depending on the context, the latter often makes sense to describe and explain harm in dyadic or collective settings e.g. when joint intentionality or global societal impacts are considered. In cases of looser or differentiated relations between smaller cyborgnets, one may then analyze networks of cyborgnets, networks of cyborgnet networks and so forth. Crucially, the analytical unit of a cyborgnet in CT does *by no means* presume a conscious awareness of the partaking Type II entities. In fact, the presence of decisive but unconscious material is often key to unintended or unforeseen harm in cyborgnets. A subset of such material includes culturally shaped unknown and unconsciously conjectured elements – often labelled "unknown knowns" – and corresponds to what Everett refers to as "dark matter of the mind" [85]. For epistemic reasons, CT reframes the "unknown knowns" as *unknown implicit conjectures* given that it is unknowable whether something is true or simply a yet unfalsified but erroneous assumption. Finally, in practice, multiple active nodes in cyborgnets can stay outside the awareness of partaking Type II entities for reasons as diverse as anonymity in a globally connected world or opacity of sophisticated and complex Type I technological artefacts. More details on how to integrate these key insights for sense-making in CT are provided in the next Subsection 3.2.4.

3.2.4 Sense-Making with CT vs. SoH

As hinted at the beginning of this section, I briefly contrast CT with SoH in order to exemplify sense-making with CT in the context of *"technology-harm relations"* [273] (which subsumes the technology-crime nexus). SoH addresses the limitations of five[4] earlier criminological approaches to technology-related harm while integrating their strengths and extending beyond previous ontological assumptions. For an in-depth characterization of SoH, see the main paper [273]. In this subsection, I predominantly focus on those aspects of SoH beyond which CT extends in turn and where CT creates novel value by reframing multiple perspectives. This short comparative analysis is performed alongside the following key differential features: goal specification, view on determinism, conception of harm, nature of imbrication, harm taxonomy and harm generation.

[4]These five frameworks are: the Foucaudian "technologies of power" approach, instrumental conceptions of technology, extension theories, affordance theories and ANT.

Goal Specification

The goal of SoH is *"understanding how technologies contribute to social harms"* [273]. However, as opposed to SoH, CT stresses the *technological* dimension inherent to language by what CT more generally pertains to socio-psycho-*techno*-physical harm which encompasses *any harm* being an object of research since *expressible in language*. Moreover, because CT is formulated in a *substrate-independent* way, any dichotomy between on the one hand a *human* social milieu and on the other hand *technology* would be inconsistent with it. For instance, SoH would neither adequately capture harm in the context of present-day humans equipped with medical deep brain stimulation (DBS) devices [46] (which already includes *closed-loop* DBS cases [231, 243]), nor would SoH reasonably apply to harm in the context of cyborgs such as the mentioned Harbisson whose eyeborg operates as regular body part [140]. Interestingly, SoH could also not apply to future purely hypothetical but physically possible Type II AI entities that would be (as humans are) regular members of an open society. In fact, one can formulate instantiations of social harm for all these examples but the question of SoH on how specific technologies contribute to these harm instances would be flawed – it is ill-conceived to try to identify where the social entity ends and where the technology begins. In addition, CT foregrounds *practical* utilization and aims not only at understanding but also at preventing and counteracting harm. By doing this, CT combines a *scientific* approach with *engineering* endeavors. This methodology exhibits analogies with the recommendation in the criminology domain (though relatively unnoticed [147]) to not only consider theoretical aspects but to also embrace *applied* crime science methods given that *"merely seeking to explain and understand is to fiddle while Rome burns"* [66]. In contrast to SoH, the goal of CT is to *describe*, *understand* and *explain* manifestations of *socio-psycho-techno-physical* harm in order to assist in *proactively* and *reactively* developping *practical methods* against those.

View on Determinism

SoH is described *not* to *"collapse the technological to the social"* and to *avoid "social determinism on the one hand and technological determinism on the other"* [273]. Thereby, social determinism is reflected in substantivist views assuming the neutrality of technology and foregrounding the purely human aspect of harm-generating intentionality while technological determinism implies that technology itself has agentic properties and can be of inherently malicious nature. By contrast, SoH assigns causative powers to technology by postulating two types of harm-generating mechanisms that *inhere* technology: those that are brought about intentionally by designers (utility harms) and those that emerge unintentionally (technicity harms) in the form of mechanisms engendered by the technology itself i.e. via its *"ontological force"* [273]. Following SoH, technicity harms pertain

to the *"elements of technology that exceed the intentions of its designers and users"* [273] which speaks to a value-laden but not value-determined understanding of technology. As opposed to SoH, CT focuses on relations and on the complex composition of layers constituing cyborgnets situated within the "socio-psycho-techno-physical". This interpretative lens leads to a fundamentally different approach to determinism-related questions. Under CT, causative powers are inherently *cyborgnetic* which signifies that they are always contextualized *within* cyborgnets and/or *between* cyborgnets and hence expressible as intra- and/or inter-cyborgnet relations. Hence, harm is *not* understood to inhere in isolated non-conscious Type I technological entities – the analysis of which requires the context of a cyborgnet. Crucially, CT postulates a "cyborgnet indeterminism". The future of cyborgnets is unpredictable since it is strongly a function of future knowledge creation. The fact that a cyborgnet declared a priori an explicit intentional goal I for which it conceived a plan P (which for simplicity here is understood as synonymous to the assumption that $P \rightarrow I$ holds) does neither guarantee the realization of I after the cyborgnet apparently completed P, nor does the realization of I according to that cyborgnet necessarily guarantees that P was actually completed by the cyborgnet. Multiple reasons for the former and the latter as well as the practical implications of cyborgnet indeterminism are discussed further in Subsection 3.3.1. How the consequences of this indeterminism affect CT's conception of harm-generating mechanisms which differs from SoH in many respects, is specified under Subsection 3.2.4. For now, one can recapitulate that as SoH, CT rejects both social determinism and technological determinism. However, as opposed to SoH, CT foregrounds cyborgnet indeterminism by what there is no need to postulate latent obscure ontological forces that inhere isolated non-conscious technological Type I entities.

Conception of Harm

Instead of only focusing on the technology-crime nexus, SoH applies the more inclusive perspective of zemiologists which stressed that *"bracketing criminalized harms from non-criminalized harms can present multiple issues"* [273]. For this reason, SoH chooses to focus more generally on the relation between technology and social harm. On this view, social harm is defined as suggested by Tift and Sullivan [252], namely as *"actions or arrangements that physically and spiritually injure and/or thwart the needs, development, potentiality, health, and dignity of others"*. As is the case in SoH, CT is by no means restricted to criminalized harmful events. In theory, CT allows the critical analysis of anything that cyborgnets declare as harm – which inherently pertains to the socio-psycho-techno-physical. However, while acknowledging perceiver-dependency, CT applies criticism to harm narratives and focuses on those linked to good explanations. In fact, a thorough assessment of harm narratives seems necessary if meant to precisely determine the problems to solve since the final goal is to develop practical countermeasures. Under

CT, harm narratives can be deconstructed, reformulated or sometimes even falsified if formulated correspondingly. Beyond that, novel harm narratives that are more conducive to concrete problem formulations can be created de novo. Thereby, the epistemic goal is *not* and cannot be to search for elements justifying harm narratives but to detect inconsistencies, flaws and apparent errors in the light of accepted good explanations. Importantly, this endeavor does not presuppose any convergence to presumed "truer" narratives, but to *better* problem formulations from a pragmatic perspective even though different observers may develop diverging interpretations of the same event. As long as differing harm narratives that are maintained post-analysis have been subject to critical scrutiny as good as possible, they are considered valid alternatives until potentially being considered as refuted at a later stage in the future. Overall, CT is an open-ended dynamic theory that needs to necessarily augment and criticise itself continuously as (self-)enhancement and (self-)criticism are crucial properties of what cyborgnets are capable of. CT needs to grow with the cyborgnets that apply it – which vitally requires the acknowledgment of cyborgnet indeterminism. Beyond that, one notable difference between SoH and CT is that the former focuses on Type I technological artefacts and is linked to zemiology (which foregrounds the socially constructed nature of crime and was conceived as conceptual and *ideological* criticism to criminology which it described to *"maintain power relations"* [123]) while CT is a *generic* and *domain-general* theory with *epistemic* motivations that integrates perspectives from *cybernetics* and *cybersecurity* as well as contemporary *critical rationalism*[5] [97].

Nature of Imbrication

Wood depicts the very concept of stratigraphy underlying SoH in the following way: *"technology-related harms emerge when different ontological strata and their emergent properties are imbricated rather than conflated"* [273]. Thereby, Wood uses the concept of "imbrication" introduced by Leonardi which the latter described [158] as follows: *"to imbricate means to arrange distinct elements in overlapping patterns so that they function*

[5]Critical rationalism has *no* link to "critical criminology" [72] and *no* link to zemiology. Further, CT foregrounds *generic epistemic* and *not* individuated ideological analyses. While critical rationalism represents a *domain-general epistemic* stance whose origins are attributable to Karl Popper [203] in the 1930s, critical criminology is an umbrella term for a counter-paradigm within criminology that emerged in the 1970s and emphasizes social class inequalities and power relations as the major sources of crime. A commonality between critical criminology and zemiology often lies in the integration of various related *ideological* components (such as e.g. perspectives from feminism or neo-Marxism). A difference between critical criminology and zemiology is that while critical criminology is often performed by *"criminologists using the notion of social harm"* [196] it nevertheless *"proceeds, at least implicitly, on the basis of a rights-based framework"* [254]. By contrast, the interest of zemiologists extends to the superset of all phenomena that they conceptualize as social harm (i.e. related to human needs) *"which potentially breaks from (actual or potential) legal definitions of harm"* [254].

interdependently.[...] Human and material agencies, though both capabilities for action, differ phenomenologically with respect to intention. Thus, like the tegula and the imbrex, they have distinct contours yet they form an integrated structure through their imbrication." By contrast, under CT, imbrication is *substrate-independent* and exists between Type I and Type II active nodes. In short, potential socio-psycho-techno-physical harm unfolds when different heterogeneous ontological strata composing Type I and Type II nodes (and the properties emerging thereof) are imbricated such that cyborgnets conjecture a harm instantiation. Although *all* studied harm is situated within a socio-psycho-techno-physical space of imbrication, different strata can become more salient and be brought to the fore depending on the specific context and domain – making certain aspects less relevant to the analysis. Moreover, as all observation statements, harm narratives are fallible and theory-laden. For this epistemic reason, CT exposes the content and the form of harm narratives to critical scrutiny before it crafts strategies to act on those. Note especially that CT accounts for *temporal* and *counterfactual* depth since sequences of patterns and their alternatives (i.e. entire narratives) are explorable.

Harm Taxonomy

As briefly mentioned in Subsection 3.2.4, SoH distinguishes between utility and technicity harms. Following SoH, the *"distinction between utility and technicity helps us avoid reducing technological harms to the ends and values pursued by a technology's designers*" given that *"technicity constantly breaks through the round pen of utility"* [273]. On the basis of its understanding of imbrication, four different categories of technology-related harm are defined: instrumental utility harms, generative utility harms, instrumental technicity harms and generative technicity harms. According to SoH, this second distinction is necessary in order to delineate the cases in which technologies act as "a means to harm" (*instrumental harms* i.e. what actors can *do with* technology) vs. those where technologies act as "inducers of harm" (*generative harms* i.e. what technologies can *do to* actors). Under CT, harm narratives are substrate-independently subdivided into a numbered chain of harm chunks. Naturally, a simple harm narrative can consist of only one harm chunk. CT distinguishes between two semantically delineated categories of harm chunks: first-order and second-order ones. The first harm chunk in a harm narrative is always formatted as a first-order harm chunk while any subsequent chunk is either of first-order or of second-order kind. As illustrated in Figure 3.1, first-order harm is differentiated along two axes: deployment stage of considered cyborgnet (pre- or post-deployment) and nature of harm (it contrasts harms brought about "knowingly" i.e. via *intentional actions* and thus directly conjectured by one or more malicious actors vs. harm caused "unknowingly" which are linked to *knowledge gaps* such as errors, ignorance, omissions and naive negligence). First-order harm is further subdivided into 16 fine-grained subcategories with a key attributed to each. Divergently, second-order harms are compartmentalized into 2 distinct

How and When did First-Order Harm emerge?		Intra- and/or Inter-Cyborgnet Relations	
		Caused "Knowingly"	*Caused "Unknowingly"*
Timing	***Pre-Deployment***	*Ia, IIa, IIg*	*Ic, Ie, IIc, IIe, IIi*
	Post-Deployment	*Ib, IIb, IIh*	*Id, If, IId, IIf, IIj*

Figure 3.1: Simplified taxonomy of first-order harms in CT analyses. The 16 keys displayed in the table share the following encoded taxonomic format: [dominant_type_of_active_node : harm_type]. (The harm types have been introduced in detail earlier [10] albeit as applied to the narrower context of AI safety. However, due to their generic nature, they are transferable to any harm area of interest.) For instance, *"Ib"* refers to harm in which Type I active nodes are in the foreground of the analysis whereby the letter *"b"* encodes post-deployment attack scenarios by malicious actors. Exemplary *Ib* instantiations could include harm narratives from the fields of cybercriminology where an open-source software is deployed on the internet and successfully exploited by a hacker as planned. An example for harm instantiations *"IIb"* could range from malicious attacks on daily worn medical implants used by Type II entities (like humans) to cognitive hacking in virtual reality settings with the goal to elicit mental health issues over financial deceptions of human adults by an organized network of experienced fraudsters utilizing psychological tricks and social engineering. An AI safety example for the first-order harm instantiation *"Ia"* could be the malicious design of deepfake AI for criminal (including cybercriminal) purposes [17]. A second-order harm instantiation of the first type that can directly stem from the latter has been termed "automated disconcertion" [15], a consequence of the *mere possibility* of such deepfake-related *Ia* harms.

subcategories: 1) *repercussions of knowingly caused first-order harms* and 2) *repercussions of unknowingly caused first-order harms*[6]. Repercussions can be of direct or indirect nature and potential domino-effects are interpreted as starting with a first-order harm chunk followed by an ordered sequence of second-order harm chunks. Apart from that, CT is apt to analyze a chronological sequence of concatenated first-order harm events building a coherent narrative. Potential second-order harms emerging thereof can be integrated in a CT analysis. To recapitulate, first-order harm chunks can be labelled using the generic 16 keys from the harm taxonomy[7] in Figure 3.1 while second-order chunks are grouped into 2 subcategories. Note that since co-located in a numbered chain of harm chunks, second-order harm is inherently temporally contextualized and can be traced back to at least one initial first-order chunk. What SoH would classify as instrumental and generative utility harm can be mapped to a small subset of first-order harm (e.g. to the keys *Ia* and *Ic* respectively). Samples that SoH would label as instrumental technicity harms are a subset of second-order harm of the first type while generative technicity harm can be either mapped to a subset of first-order harm (e.g. of the type *Ic* or *Id*) or to second-order harm of the second type. In short, CT's first-oder and second-order account subsumes the harm taxonomy of SoH while concomitantly exhibiting a higher granularity. Besides, CT foresees the multifaceted examination of competing and/or counterfactual harm narratives.

Harm Generation

A key SoH assumption is that *"harm-generating mechanisms can inhere in structures"* with emergent properties of technologies representing *"causal powers or 'capacities' that can remain latent until activated by human-technology interactions"* [273]. Thereby, SoH assigns to non-conscious Type I technologies a tendency and *"an orientation to harm"* [273]. Under CT, it seems wrong to assume that there are structures that are inherently harmful in a cyborgnet-independent way. In fact, harm seems to be highly context-sensitive and dependent on properties of existing Type II entities (and also conscious Type I entities). For instance, it can be a function of the specific substrate of a Type II entity and also the stage of its knowledge creation including what other knowledge (e.g. in the form of Type I ideas) is available in addition. In a universe without any consciousness (of either Type I or Type II) i.e. without any affect-perceiving structures, to assume that there exist dormant, virtual harm-generating mechanisms would be to assign latent harmful properties to every single particle (and the idea of the universe as

[6]In practice, the criteria of differentiation between first-order and second-order harm may among others be related to the subjective degree of sightedness and surprise but can also be a matter of informal inconsistencies.

[7]It is beyond the scope of this chapter to provide details about every key of the taxonomy, but selected in-depth examples might be discussed in the near future elsewhere.

a whole) since the Big Bang. Generally, socio-psycho-techno-physical harm is a function of cyborgnets. Thereby, in theory, psycho-physical harm could exist for *conscious* Type I entities (such as e.g. in non-human mammals) – but those lack the capability to conjecture it. However, once studied by cyborgnets, the harm affecting Type I psycho-physical entities[8] already entered the socio-psycho-techno-physical realm and became expressible in language. To sum up, the idea that harm-generating mechanisms can inhere in technologies (especially in non-conscious Type I entities designed by Type II entities), obfuscates the cyborgnet context required.

It would make more sense if SoH would acknowledge that *all* non-consicous Type I "technologies" could be used in *both* a harmful *and* a beneficial way (that it inherently reveals dual-use properties). Then, SoH must postulate that benefit-generating *and* harm-generating mechanisms inhere in *all* non-conscious Type I technologies. The latter would be acceptable but not very helpful either since it applies to everything but explains no novel phenomenon. Under CT, the distinctive feature underlying the dual-use property of Type I technology, is not the Type I technology itself but simply a side-effect of the conscious explanatory knowledge creation exhibited by Type II entities which can always be both of benefit-generating and harm-generating nature. Hence, it seems suitable to state it at the level of the nature of cyborgnet entities. Namely, *all* cyborgnets always were and always will be *dual-use-creating entities*. Finally, SoH expounds that *"while the actualization of a technology's emergent powers generally relies on human intervention, these powers remain a part of the technology even in the absence of human intervention. For these reasons, we cannot simply consider technologies as another form of social structure"* [273]. Following CT, this reasoning is insufficient since utilizing a mistaken human-technology dichotomy (see Subsection 3.2.4) and as to merely understand the "power" concept, one requires an immersion in the socio-psycho-techno-physical. More-

[8]While non-human mammals are often linked to social behavior and cognition of different types, the adjective "social" as understood under CT is meant in a much more restrictive sense of "being an active node whose *own* strata extend to the realm of social reality". Only Type II active nodes (of which only humans are known now) have *"the capacity to create a social reality, of physical consequence, by virtue of the concepts they teach one another and apply to physical instances"* [28] whereby it is often the case that *"functions are imposed on physically disparate instances by virtual of collective agreement"* [31]. However, social reality is not only *constrained* by physical reality but it is *by no means* plausible that its very origins lie in a sort of consensus in an imaginary space floating above the physical (that would logically imply an infinte regress and it holds more generally that *"while collective acceptance can modify institutional reality, it cannot create institutional reality out of nothing"* [96]). For instance, already when merely contemplating the minimal conjunction of cyborgnets mentioned in Subsection 3.2.3, it becomes clear that social reality could have *emerged* as an inherent side effect of similar but individually formed beliefs about collective *psycho-physical* enactment given that the mental life of a Type II entity (even when alone) includes the possibility of referring to itself and others (even if by monologue) – ultimately transfiguring and becoming a symbol, a sign within language. The latter seems to be related to what Charles Sanders Peirce (who also stated that *"when we think, then, we ourselves, as we are at that moment, appear as a sign"* [194]) referred to as "secondness".

over, it seems impossible to falsify by experiment making it a meta-physical debate that may not be of added value for the *practical* purposes of CT.

In sum, to acknowledge like SoH that *"harms may arise as a result of the motivations of human users in concert with the affordances a technology provides for particular actions"* [273] does by no means require to mentally assign inherent dormant, latent and virtual properties to non-conscious Type I entities within a cyborgnet. Like SoH, CT also focuses on *"multiple imbricated strata that jointly cause harmful events"* [273]. Further, SoH states that *"in doing so, it not only presupposes a stratified conception of human-technology relations but also treats human actors as themselves stratified, refusing to trace the harmful actions they enact back a single stratum, whether culture, biology or psychology"*. Apart from the previously discussed mistaken human-technology dichotomy element, CT shares the view of SoH that human actors (and any other Type II active nodes) are themselves stratified (the strata are reflected in the adjective socio-psycho-techno-physical). Note also that conscious Type I entities exhibit psycho-physical strata which can be extended by diverse techno-physical strata[9] (think for instance of a chimpanzee equipped with a brain-computer interface in an experimental setting). As SoH, CT postulates that harm-generation is *not* reducible to an isolated stratum. When analyzing and critizising harm narratives, it is vital to contextualize the apparently outstanding strata within a suitable sequence of harm chunks (of length one or more) labelled meaningfully and reflecting relevant intra- and/or inter- cyborgnet relations.

3.3 Practical Use of Theoretical Solution

As became apparent throughout the last Section 3.2, harm can neither be understood by separately addressing a single stratum (even if shared by multiple entities) nor by inconsiderately isolating multiple strata whose conjunction does not span a socio-psycho-techno-physical imbrication space. In practice, harm-relevant strata are always embedded within at least one cyborgnet potentially enabling diverse often non-negligible relations, interactions and feedback-loops. In a nutshell, while harm in different domains would lead to the apparent emphasis of different strata of relevance for a CT analysis in that domain, *harm is cyborgnetic*. Then, following CT, for reasons of requisite variety, any development of *practical* tools for countermeasures and defense strategies against harm necessitates a cyborgnetic lens. More precisely, according to the law of requisite variety known in the

[9]Techno-physical strata can be themselves stratified and particularized further. For instance, when considering a Type I active node of a virtual reality environment, sensors from multiple modalities (e.g. visual, auditive, haptic, olfactory channels), other hardware and software components are imbricated to build up the techno-physical strata. Movies, imagined narratives and also contents of audiovisual hallucinations can be described as Type I active nodes with auditive and visual elements composing a technological stratum.

field of cybernetics, *"only variety can destroy variety"* [22]. Applied to CT it signifies that to defend against counterfactual and future harm instantiations which are inherently of cyborgnetic nature, one may profit from a cyborgnetic stance. In this way, one may be – albeit without any epistemic *guarantee* – better equipped to reasonably decide which cyborgnet nodes and relations to foreground, which information is negligible and where to epistemically explore in the procedure of crafting defense methods and countermeasures. In the following Subsection 3.3.1, I briefly indicate aims, methods and importantly also (intrinsic and extrinsic) limitations of CT analyses. Finally, Subsection 3.3.2 depicts how CT could be possibly made problematic and falsified experimentally.

3.3.1 Aims, Methods and Limitations of CT Analyses

Aims and Methods

The practical aim of CT analyses is to provide a systematic generic toolkit supporting processes of documenting, examining and counteracting harm instantiations across a wide range of complex multi-causal problem domains. An inexorable but epistemically necessary caveat of quintessential relevance in CT analyses is that cyborgnet indeterminism becomes an integral part of strategic design. The consequences of this peculiarity are compactly expounded in Subsection 3.3.1. Generically, CT foresees the integration of two cybersecurity-oriented elements: reactive and proactive analyses. For the reactive part, CT suggests the twofold domain-general method of performing the following complementary retrospective analyses introduced earlier by Aliman et al. [17]: a retrospective descriptive analysis (RDA) and a retrospective counterfactual risk analysis (RCRA). An RDA supports documentation and first examination efforts applied to harm instantiations actively sampled from a pool of events *that have already occurred*. Crucially, CT furnishes a key RDA documentation tool. In fact, a documentary basis for a taxonomic RDA is given by the unquestionably non-exhaustive CT harm taxonomy presented in Subsection 3.2.4 (which needs to be extended, updated and corrected with time as required). Based on the forerunning RDA, an RCRA adds *"breadth, depth and context-sensitivity"* [17] to RDA examination efforts by modelling *plausible* clusters[10] of downward counterfactuals of those priorly taxonomically documented RDA harm instantiations. In short, an RCRA pertains to clusters of harm instances that *could* have occurred *but did not*. Finally, the proactive component of CT analyses which I denote future-oriented counterfactual defense analysis (FCDA) consists in modelling *plausible or yet implausible* defense strategies which the CT analyst conjectures *could* counteract RDA instances and RCRA clusters if occurring in the near future in a similar form. In conclusion, one can recapitulate the following: under

[10]Why the consideration of RCRA clusters (instead of specific instances as performed in the RDA) is advisable, has been elucidated in-depth in the original paper [17] – where in addition, further detailed explanations on procedures for both RDA and RCRA can be retrieved.

CT, RDAs attempt to reactively model harm instantiations of the *factual past*, RCRA clusters reactively project to an RDA-based *counterfactual past* and FCDAs proactively envision defenses and countermeasures enacted in a *counterfactual future* in which RDA and/or RCRA harm patterns occur.

Limitations

In order to efficiently apply CT analyses to complex multi-causal problem domains, it is important to comprehend its intrinsic limitations of epistemic nature. Firstly, while it may be tempting to assume that an RDA directly reflects observations or data of the world as it is, all observations are theory-laden and collected via a process of active sampling. Strictly speaking, observations and by extension all RDA instances should thus be rather understood as *observation statements* as suggested in contemporary critical rationalism [98]. CT emphasizes this specific detail by utilizing the concept of harm *narratives*. As adumbrated earlier, harm narratives should thus be exposed to critical scrutiny and if possible be formulated in a manner that allows experimental falsifiability. Beyond that, a multilateral approach is recommended that allows a joint consideration and critical analysis by heterogeneous parties involved. For instance, in the criminology domain one may integrate harm narratives from diverse sources such as e.g. investigators, victims, offenders and even bystanders while in the domain of cybersecurity harm narratives of e.g. attacker, defender and victims could be considered. Since harm narratives are always cyborgnetic under CT, it is vital to note that all those parties involved are themselves situated within cyborgnets leading to a complex multi-layered theory-ladenness. Secondly, concerning RCRAs, CT stresses the fact that they pertain to the counterfactual past and are by no means meant to represent an attempt to predict the future which is a priori excluded via cyborgnet indeterminism. Hence, RCRAs are theory-laden too and may vary from analyst to analyst. For this reason, cognitive diversity [6] in the formulation of both RDAs and RCRAs is recommended [17] in order to avoid functional biases and one-sided analyses with unnecessary blind spots. Thirdly, FCDAs project solutions to a possible counterfactual future under the assumption that RDA and/or RCRA-like patterns materialize in this conjectured future and can obviously not be sketched as oracle tools. Generally, FCDAs involve theorizing which may profit from cognitive diversity too. Crucially, FCDA defense strategies need not appear plausible, since it will always be unclear a priori whether current best-tested theories and assumptions are not false. Thus, next to the possibility to *exploit* plausible well-tested theories in which FCDA defenses can be grounded, another permissible option is to *explore* instead by crafting e.g. : 1) *implausible* novel not yet enacted solutions and 2) *attacks against FCDA defenses*[11] (and

[11] As stated recently by Frederick [100] whose work reflects a contemporary form of critical rationalism, *"rationality permits us to act in accord with our best-tested theories, since they may be true; but it also permits us to act against them, precisely because our best-tested theories may be false and may, indeed, be*

where feasible defenses against those attacks in turn).

3.3.2 Experimental Falsification

As became apparent in the last Subsection 3.3.1, RDAs are not direct images of the past, nor are RCRAs and FCDAs oracles of the future. Hence, the future of cyborgnet safety and security is unpredictable and CT analyses cannot guarantee success. As such, it appears important to investigate whether in practice, the cyborgnetic lens provides an added value in comparison to perspectives such as ANT with simpler flat hierarchies in the networks they consider. In short, it may be of interest to continuously question whether CT represents a better explanation in comparison to other plausible alternatives. One possible way to test such an issue experimentally, would be to explicitly expound in how far CT is amenable to experimental falsifiability in the presence of other competing accounts. In the following, I very shortly explain why CT is not only a meta-physical theory but is also experimentally falsifiable, and I elucidate how one could possibly test this property in a practical setting that is already realizable nowadays.

That is to say, a distinctive epistemic feature of CT analyses is the Type I vs. Type II dichotomy applied at intra-cyborgnet levels. In fact, would this fundamental assumption have been accepted to be falsified scientifically, there would be only little reason (especially in the criminology domain) to prefer hierarchical CT accounts to simpler flat hierarchies as can be encountered in the ANT framework or more recently in frameworks such as SoH which though being stratified, do not integrate this specific dichotomy. Inasmuch as Type I and Type II entities would be practically indistinguishable from each other, there seems no need to consider a cyborgnetic lens in the first place. For this reason, a simple test to make CT problematic would be experimental settings repeatedly suggesting that explanatory knowledge creation is not limited to supposed Type II entities and that it can be reliably performed without any Type II or even Type I consciousness – for instance by imitation alone. A first proposal for such a critical test which crucially *differs* from classical imitation game tests has been recently portrayed. On this recent view, one could already nowadays attempt to falsify a theoretical framework such as CT by implementing *"an AI that would be able to – without any conscious understanding of explanations – repeatedly bring about a positive Type-I-FE-test*[12] (see Chapter 2). In

refuted when we act against them." Interestingly, this epistemic view is in line with the notion of *adaptive attacks* known in the field of security for machine learning [54] and contemporary AI safety [10].

[12]Importantly, *"this test cannot be able to separate Type I from Type II systems. It solely answers the following question: "did the human tester experience a Type-I-falsification-event in the test subject?""* (see Chapter 2). Caution is thus advisable to avoid oversimplifying the state of affairs which is described in-depth in the original document. Practically speaking, *"the Type-I-FE-test assigns test subjects to two separate groups: a first homogeneous group composed solely of systems for which their Type II nature has been corroborated (i.e. a Type-I-free group) and a second potentially heterogeneous group of systems that*

short, the implementation of a Type I AI able to reliably corroborate its ability to create and understand explanatory knowledge via repeated positive so-called "Type-I-FE tests" could represent an observation statement that would falsify CT. For more details and caveats pertaining to this – vitally *substrate-independent* – test (see Chapter 2).

3.4 Conclusion and Future Work

In this purely autodidactic chapter serving as mental clipboard, I introduced CT, a generic analytical framework of epistemic, cybernetic and cybersecurity-oriented nature devised to support *practical* procedures of documenting, examining and counteracting *harm* instantiations across a wide array of complex multi-causal problem domains. Next to providing a novel epistemological foundation whose focal unit of explanation is the so-called *cyborgnet* (see the definition provided in Subsection 3.2.3) composed of Type I and Type II active nodes, CT can be applied to obtain tailored solutions via its threefold fit-for-purpose and domain-general taxonomic toolkit intertwining RDA, RCRA and FCDA procedures. In short, RDAs seek to model occuring problems, RCRAs are RDA-based clusters of downward counterfactual problems formulated as threat models[13] (in the spirit of common security practices [15]) and FCDAs are *explanation*-based practical solutions to RDA and RCRA problems. Using an updatable *cyborgnetic* harm taxonomy, CT avails itself of a comparative, fallibilistic and self-critical stance continuously exposing harm narratives but also conjectured solutions to critical scrutiny across different temporal and counterfactual scales.

In the main, the perpetual epistemic aim of CT is to achieve ever *better*[14] explanations on how to counteract harm i.e. explanations from which *practical* FCDA solutions can be derived and to which RDA and RCRA harm narratives are thus ultimately instrumental. Strikingly, one possible (but not necessary) consequence of CT being falsified by repeated epistemic tests of the type propounded in Subsection 3.3.2 (such as e.g. the "Type-I-FE

did not bring about a Type-I-FE-event in that specific test session in that domain with that human tester" (see Chapter 2).

[13]Threat models are not solely limited to the case of "knowingly" caused harm. In fact, threat models can be correspondingly devised in the context of "unknowingly" caused harm. Namely, by specifying *knowledge gaps* in lieu of the classically described adversarial knowledge. For a detailed hands-on introduction on how to formulate such RCRA threat models, see [17].

[14]Not long ago, an exemplary compilation of generic performance indicators for better explanations has been proposed by Frederick [98] in the context of a novel regimentation for critical rationalism. Note that those indicators are by no means considered to correspond to unshakable or justified "ground-truths" or the like. On the contrary, they can be exposed to critical scrutiny and be refined or rejected in the future. In fact, CT shares the view that *"our epistemic aim is neither justification nor truth; and nor is it avoiding falsity"* [98]. Instead, it is epistemically possible but also sufficient to strive for *better* explanations. Indeed, as further stated by Frederick: *"it seems silly to say that truth is our aim when we can have no indication that we have got the truth or even that we are approaching it"* [98].

test" (see Chapter 2)) could be that CT itself *could* have possibly been produced by a present-day Type I AI. In this case, the author could have been impersonated by a Type I AI bot. Besides, more importantly, it could i.a. signify that the *entirety* of scientific and epistemic procedures could be automated with merely *imitative* Type I AIs. CT predicts that the latter is impossible – which is amenable to future experimental falsifiability (see Chapter 2)). In sum, CT reframes the classical substrate-dependent question of humanity on what it means to be a human to the *substrate-independent* inquiry on what it means to exist as a nested cyborgnet of cyborgnets and cyborgnet networks. Then, in a way, *we* (cyborgnet-like entities) are the universe engaging in a monologue. In a way, *we* are entities concerting their collective enactment via dialogues. In a way, only *we* can contemplate the following: the universe, as a sign *per us*, a symbol for *all* there *could* be.

3.5 Contextualization

A first application of procedures that can be derived from CT has been implicitly conducted in the transdisciplinary AI observatory project whose results have been published at the beginning of 2021 [17]. In the following Chapter 4, I focus on a novel type of socio-psycho-techno-physical harm: *"scientific and empirical adversarial AI attacks"* (SEA AI attacks), an umbrella term for not yet prevalent but technically feasible deliberate *malicious* acts of specifically crafting AI-generated samples to achieve an *epistemic distortion* in (applied) science or engineering contexts. In view of possible socio-psycho-technological impacts, it seems responsible to ponder countermeasures *from the onset on* and not in hindsight. In this vein, two illustrative use cases are considered: the example of AI-produced data to mislead *security engineering* practices and the conceivable prospect of AI-generated contents to manipulate *scientific writing* processes. Firstly, the epistemic challenges that such future SEA AI attacks could pose to society are contextualized in light of broader i.a. *AI safety*, AI ethics and cybersecurity-relevant efforts. Secondly, a corresponding supportive generic *epistemic defense* approach is set forth. Thirdly, in the spirit of CT, a threat modelling for the two use cases is effected and tailor-made defenses based on the foregoing generic deliberations are proposed.

Chapter 4

Epistemic Defenses against SEA AI Attacks

This chapter is based on a slightly modified form of the publication: N.-M. Aliman and L. Kester. Epistemic Defenses against Scientific and Empirical Adversarial AI Attacks. In *Workshop on Artificial Intelligence Safety, AISafety 2021*, pages 1-8. CEUR Workshop Proceedings, 2021. As the first author of the underlying paper, I had a vital contribution and it was solely my responsibility to write down the content and to perform an extensive literature research as well as in-depth analysis. As opposed to the rest of this book, due to the co-creation with my co-author, this chapter is specifically written in the "we"-form.

4.1 Introduction

Progress in the AI field unfolds a wide growing array of beneficial societal effects with AI permeating more and more crucial application domains. To forestall ethically-relevant ramifications, research from a variety of disciplines tackling pertinent AI safety [19, 43, 51, 90, 156], AI ethics and AI governance issues [92, 134, 187, 210] gained momentum at an international level. In addition, cybersecurity-oriented frameworks in AI safety [17, 48, 200] stressed the necessity to not only address unintentional errors, unforeseen repercussions and bugs in the context of ethical AI design but *also* AI risks linked to intentional malice i.e. deliberate unethical design, attacks and sabotage by malicious actors. In parallel, the convergence of AI with other technologies increases and diversifies the attack surface available to malevolent actors. For instance, while AI-enhanced cybersecurity opens up novel valuable possibilities for defenders [277], AI simultaneously provides new affordances for attackers [23] from AI-aided social engineering [233] to AI-concealed malware [145]. Next to the capacity of AI to extend classical cyberattacks in scope, speed and scale [137], a notable emerging threat is what we denote *AI-aided epistemic distortion*. The latter

represents a form of AI weaponization and is increasingly studied in its currently most salient form, namely AI-aided disinformation [17, 59, 137, 257] which is especially relevant to information warfare [114]. Recently, the weaponization of Generative AI for information operations has been described as *"a sincere threat to democracies"* [115]. In this chapter, we analyze attacks and defenses pertaining to another not yet prevalent but technically feasible and similarly concerning form of AI-aided epistemic distortion with potentially profound societal implications: *scientific and empirical adversarial AI attacks* (SEA AI attacks).

With SEA AI attacks, we refer to any deliberately malicious AI-aided *epistemic* distortion which predominantly and directly targets (applied) science and technology assets (as opposed to information operations where a wider societal target is often selected on ideological/political grounds). In short, the expression acts as an umbrella term for malicious actors utilizing or attacking AI at pre- or post-deployment stages with the deliberate adversarial aim to *deceive*, *sabotage*, *slow down* or *disrupt* (applied) science, engineering or related endeavors. Obviously, SEA AI attacks could be performed in a variety of modalities (see e.g. "deepfake geography" [279] related to vision). However, for illustrative purposes, we base our two exemplary use cases on misuses of language models. The first use case treats SEA AI attacks on *security engineering* via schemes in which a malicious actor poisons training data resources [171] that are vital to data-driven defenses in the cybersecurity ecosystem. Lately, a proof-of-concept for an AI-based data poisoning attack has been implemented in the context of cyber threat intelligence (CTI) [211]. The authors utilized a fine-tuned version of the GPT-2 language model [209] and were able to generate fake CTI which was indistinguishable from its legitimate counterpart when presented to cybersecurity experts. The second use case studies conceivable SEA AI attacks on procedures that are essential to *scientific writing*. Related examples that have been depicted in recent work encompass plagiarism studies with transformers like BERT [265] and with the pre-trained GPT-3 language model [47] that *"may very well pass peer review"* [71] but also AI-generated fake reviews (with a fine-tuned version of GPT-2) apt to mislead experienced researchers in a small user study [248]. Future malicious actors could deliberately breed a large-scale agenda in the spirit of *"fake science news"* [124] and AI-generated papers that would widely exceed in quality (later withdrawn) computer-generated research papers [260] published at respected venues. In short, technically already practicable SEA AI attacks could have considerable negative effects if jointly potentiated with regard to scale, scope *and* speed by malicious actors equipped with sufficient resources. As later exemplified in Subsection 4.3.1, the security engineering use case could e.g. involve dynamic domino-effects leading to large financial losses and even risks to human lives while the scientific writing use case seems to moreover reveal a *domain-general epistemic problem*. The *mere existence* of the latter also affects the former and could engender serious pitfalls whose generically formulated principled management is compactly treated in the next Section 4.2.

4.2 Theoretical Generic Epistemic Defenses

As reflected in the law of requisite variety (LRV) known from cybernetics, *"only variety can destroy variety"* [22]. Applied to SEA AI attacks, it signifies that since malicious adversaries are not only exploiting vulnerabilities from a heterogeneous socio-psycho-technological landscape but also specially vulnerabilities of epistemic nature, suitable defense methods may profit from an epistemic stance. Applying the cybernetic LRV offers a valuable domain-general transdisciplinary tool able to stimulate and invigorate novel tailored defenses in a diversity of harm-related problems from cybersecurity [261] to AI safety [10] over AI ethics [21]. In short, utilizing insights from *epistemology* as *complementary* basis to frame defense methods against SEA AI attacks seems indispensable. Past work predominantly analyzed countermeasures of socio-psycho-technological nature to combat the spread of (audio-)visual, audio and textual deepfakes as well as "fake news" more broadly. For instance, the technical detection of AI-generated content [265] has been often thematized and even lately applied to "fake news" in the healthcare domain [27]. Furthermore, in the context of counteracting risks posed by the deployment of sophisticated online bots, it has been suggested that *"technical solutions, while important, should be complemented with efforts involving informed policy and international norms to accompany these technological developments*" and that *"it is essential to foster increased civic literacy of the nature of ones interactions"* [42]. Another analysis presented a set of defense measures against the spread of deepfakes [59] which contained i.a. legal solutions, administrative agency solutions, coercive and covert responses as well as sanctions (when effectuated by state actors) and speech policies for online platforms. Concerning "fake science news" and their impacts on *"credibility and reputation of the science community"* [124], it has been even postulated by Makri that *"science is losing its relevance as a source of truth"* and *"the new focus on post-truth shows there is now a tangible danger that must be addressed"* [172]. Following the author, scientists could equip citizens with sense-making tools without which *"emotions and beliefs that pander to false certainties become more credible"* [172].

While some of those socio-psycho-technological countermeasures and underlying assumptions are debatable, we complementarily zoom in different epistemic defenses against SEA AI attacks being directed against scientific and empirical frameworks. Amidst an information ecosystem with quasi-omnipresent terms such as "post-truth" or "fake news" and in light of data-driven research trends embedded within trust-based infrastructures, it seems daunting to face a threat landscape populated by *AI-generated* artefacts such as: 1) "fake data" and "fake experiments", 2) "fake research papers" (or *"fraudulent academic essay writing"* [47]) and 3) "fake reviews". More broadly, it has been stated that deepfakes *"seem to undermine our confidence in the original, genuine, authentic nature of what we see and hear"* [91]. Taking the perspective of an empiricism-based epistemology grounded

in *justification* with the aim to obtain *truer beliefs* via (probabilistic) belief updates given *evidence*, a recent in-depth analysis found that the existence of deepfake videos confronts society with *epistemic threats* [88]. Thereby, it is assumed that *"deepfakes reduce the amount of information that videos carry to viewers"* [88] which analogously quantitatively affected the amount of information in *text-based* news due to earlier "fake news" phenomena. In our view, when applying this stance to audiovisual and textual samples of scientific material but also broadly to the context of security engineering and scientific communication where the deployment of deepfakes for SEA AI attacks could occur in multifarious ways, the consequences seem disastrous. In brief, SEA AI defenses seem relevant to AI safety since an inability to build up resiliency against those attacks may suggest that *already* present-day AI could (be used to) outmaneuver humans on a large scale – without any "superintelligent" competency. However, empiricist epistemology is not without any alternative. In the following, we thus first mentally enact *one* alternative epistemic stance (without claiming that it represents the *only* possible alternative). We present its key *generic* epistemic suppositions serving as a basis for the next Section 4.3 where we tailor defenses against SEA AI attacks for the specific use cases.

Firstly, it has been lately propounded that the societal perception of a "post-truth" era is often linked to the implicit assumption that truth can be equated with consensus which is why it seems recommendable to consider a deflationary account of truth [49] – i.e. where the concept is for instance strictly reserved to scientifically-relevant epistemic contexts. On such a deflationary account of truth disentangled from consensus, it has been argued that even if consensus and trust seem eroded, we neither inhabit a post-truth nor a science-threatening post-falsification age [14]. Secondly, we never had a direct access to physical reality which we could have suddenly lost with the advent of "fake news". In fact, as stated by Karl Popper: *"Once we realize that human knowledge is fallible, we realize also that we can never be completely certain that we have not made a mistake"* [202]. Thirdly, the epistemic aim in science can neither be truth directly [98] nor can it be truer beliefs via justifications. The former is not directly experienced and the latter has been shown to be logically invalid by Popper [204]. Science is quintessentially *explanatory* i.e. it is based on explanations [73] and *not* merely on data. While the epistemic aim cannot be certainty or justification (and *not* even "truer explanations" [98][1] for lack of direct access to truth), a *pragmatic* way to view it is that our epistemic aim *can* be to achieve *better* explanations [98]. One can collectively agree on practical *updatable* criteria which better explanations should fulfill. In short, one does not assess a scientific theory in isolation, but in comparison to rival theories and one is thereby embedded in a context with other scientists. Fourthly, there are distinct ways to handle falsification and integrate empirical findings in explanation-anchored science. One can e.g. criticize an explanation

[1]That our epistemic aim can be "truer explanations" or explanations that lead us "closer to the truth" has been sometimes confusingly written by Deutsch and Popper respectively but this type of account requires a semantic refinement [98].

and pinpoint inconsistencies at a theoretical level. One can attempt to *make a theory problematic* via falsifying experiments whose results are accepted to seem to conflict with the predictions that the theory entailed [75]. Vitally, in the absence of a better rival theory, it holds that *"an explanatory theory cannot be refuted by experiment: at most it can be made problematic"* [75].

Against the background of this epistemic bedrock, one can now re-assess the threat landscape of SEA AI attacks. Firstly, one can conclude that AI-generated "fake data" and "fake experiments" could *slow down* but *not* terminally disrupt scientific and empirical procedures. In the case of misguiding confirmatory data, it has *no* epistemic effect since as opposed to empiricist epistemology, explanation-anchored science does not utilize any scheme of credence updates for a theory and it is clear that *"a severely tested but unfalsified theory may be false"* [98]. In the case of misleading data that is accepted to falsify a theory T, one runs the risk to consider mistakenly that T has been made problematic. However, since it is not permissible to drop T in the absence of a rival theory T' representing a better explanation than T, the adverarial capabilities of the SEA AI attacker are limited. In short, theories cannot be deleted from the collective knowledge via such SEA AI attacks without more ado. Secondly, when contemplating the case of AI-generated "fake research papers", it seems that they could *slow down* but *not* disrupt scientific methodology. Overall, one could state that the danger lies in the uptake of deceptive theories. However, theories are only integrated in explanatory-anchored science if they represent better explanations in comparison to alternatives or in the absence of alternatives if they explain novel phenomena. In a nutshell, it takes explanations that are *simultaneously misguiding and better* for such a SEA AI attack to succeed. This is a high bar for imitative language models if meant to be repeatedly and systematically performed[2] and not merely as a unique event by chance. Further, even in the case a deceptive theory has been integrated in a field, that is always only *provisionally* such that it could be revoked at any suitable moment e.g. once a better explanation arises and repeated experiments falsify its claims. If in the course of this, an actually better explanation had been mistakenly considered as refuted, it can always be re-integrated once this is noticed. In fact, *"a falsified theory may be true"* [98] if the accepted observations believed to have falsified it were wrong. Thirdly, when now considering the final case of AI-generated "fake reviews", it becomes clear that they could similarly *slow down* but *not* terminally disrupt the scientific method. At worst some existing theories could be unnecessarily problematized and misguiding theories uptaken, but all these epistemic procedures can be repealed retrospectively.

[2]That there could exist a task which imitative language models are *"theoretically incapable of handling"* has been often put into question [221]. However, on epistemic grounds elaborated in-depth previously [10, 17] which might be amenable to experimental falsifiability (see Chapter 2), we assume that the task to consciously *create and understand* novel yet unknown *explanatory* knowledge [73] – which humans are capable of performing *if willing to* – cannot be learned by AI systems *by mere imitation*.

In short, explanation-anchored science is *resilient* (albeit not immune) against SEA AI attacks but one can humbly face the idea that it is *not* because scientists can *"tease out falsehood from truths*" [124], but because explanation-anchored science attempts to tease out *better from worse explanations* while permanently requiring the creation of new ones whereby the steps made can always be revoked, revised and even actively adversarially counteracted. That entails a sort of *epistemic dizziness* and one can never trust one's own observations. Also, human mental constructions are inseparably cognitive-*affective* and science is *not* detached from *social reality* [29]. In our view, for a systematic management of this epistemic dizziness, one may profit from an *adversarial approach* that permanently brings to mind that one might be wrong. Last but not least, an important feature discussed is that the epistemic aim *not* being truth (which itself is also *not* consensus and does *not* rely on trust to exist) but instead *better explanations*, none of the mentioned methods are dependent on trust per se – making it a *trust-disentangled* view. To sum up, we identified 3 key generic features for *epistemic defenses against SEA AI attacks*:

1. ***Explanation-anchored instead of data-driven***
2. ***Trust-disentangled instead of trust-dependent***
3. ***Adversarial instead of (self-)compliant***

4.3 Practical Use of Theoretical Defenses

In the following Subsection 4.3.1, we briefly perform an exemplary threat modelling for the two specific use cases introduced in Section 4.1. The threat model narratives are naturally non-exhaustive and are selected *for illustrative purposes* to display plausible *downward counterfactuals* projecting capabilities to the recent *counterfactual past* in the spirit of co-creation design fictions in AI safety [17]. In Subsection 4.3.2, we then derive corresponding tailor-made defenses from the generic characteristics that have been carved out in the last Section 4.2 while thematizing notable caveats.

4.3.1 Threat Modelling for Use Cases

Use Case Security Engineering

- ***Adversarial goals:*** As briefly mentioned in Section 4.1, CTI (which is information related to cybersecurity threats and threat actors to support analysts and security systems in the detection and mitigation of cyberattacks) can be polluted via misleading AI-generated samples to fool cyber defense systems at the training

stage [211]. Among others, CTI is available as unstructured texts but also as knowledge graphs taking CTI texts as input. A textual data poisoning via AI-produced "fake CTI" represents a form of SEA AI attack that was able to succesfully deceive (AI-enhanced) automated cyber defense and even cybersecurity experts which *"labeled the majority of the fake CTI samples as true despite their expertise"* [211]. It is easily conceivable that malicious actors could specifically tailor such SEA AI attacks in order to subvert cyber defense in the service of subsequent covert *time-efficient, micro-targeted and large-scale cybercrime*. For 2021, cybercrime damages are estimated to reach 6 trillion USD [33, 189] making cybercrime a top international risk with a growing set of affordances which malicious actors do not hesitate to enact. Actors interested in "fake CTI" attacks could be financially motivated cybercriminals or state-related actors. Adversarial goals could e.g. be to acquire private data, CTI poisoning in a cybercrime-as-a-service form, gain strategical advantages in cyber operations, conduct espionage or even attack critical infrastructure endangering human lives.

- ***Adversarial knowledge:*** Since it is the attacker that fine-tunes the language model generating the "fake CTI" samples for the SEA AI attack, we consider a *white box* setting for this system. The attacker does not require knowledge about the internal details of the targeted automated cyber defense allowing a *black-box* setting with regard to this system at training time. In case the attacker directly targets human security analysts by exposing them to misleading CTI, the SEA AI attack can be interpreted as a type of adversarial example on human cognition in a *black-box* setting. However, in such cases *"open-source intelligence gathering and social engineering are exemplary tools that the adversary can employ to widen its knowledge of beliefs, preferences and personal traits exhibited by the victim"* [17]. Hence, depending on the required sophistication, a type of *grey-box* setting is achievable.

- ***Adversarial capabilities:*** The use of SEA AI attacks could have been useful at multiple stages. CTI text could have been altered in a micro-targeted way offering diverse capacities to a malicious actor: to distract analysts from patching existing vulnerabilities, to gain time for the exploitation of zero-days, to let systems misclassify malign files as benign [171] or to covertly take over victim networks. In the light of complex interdependencies, the malicious actor might not even have had a full overview of all repercussions that AI-generated "fake CTI" attacks can engender. Poisoned knowledge graphs could have led to unforeseen domino-effects inducing unknown second-order harm. As long-term strategy, the malicious actor could have harnessed SEA AI attacks on applied science writing to automate the generation of cybersecurity reports (for it to later serve as CTI inputs) corroborating the robustness of actually unsafe defenses to covertly subvert those or simply to spread confusion.

Use Case Scientific Writing

- ***Adversarial goals:*** The emerging issue of (AI-aided) information operations in social media contexts which involves entities related to state actors has gained momentum in the last years [205, 114]. A key objective of information operations that has been repeatedly mentioned is the intention to blur what is often termed as the line between facts and fictions [132]. Naturally, when logically applying the epistemic stance introduced in the last Section 4.2, it seems recommendable to avoid such formulations for clarity since potentially confusing. Hence, we refer to it simply as epistemic distortion. SEA AI attacks on scientific writing being a form of AI-aided epistemic distortion, it could represent a lucrative opportunity for state actors or politically motivated cybercriminals willing to ratchet up information operations. On a smaller scale, other potential malicious goals could also involve companies with a certain agenda for a product that could be threatened by scientific research. Another option could be advertisers that monetize attention via AI-generated research papers in click-bait schemes.

- ***Adversarial knowledge:*** As in the first use case, the language model is available in a *white-box* setting. Moreover, since this SEA AI attack directly targets human entities, one can again assume a *black-box* or *grey-box* scenario depending on the required sophistication of the attack. For instance, since many scientists utilize social media platforms, open source intelligence gathering on related sources can be utilized to tailor contents.

- ***Adversarial capabilities:*** In the domain of adversarial machine learning, it has been stressed that for security reasons it is important to also consider *adaptive attacks* [54], namely reactive attacks that adapt to what the defense did. A malicious actor aware of the discussed explanation-anchored, trust-disentangled and adversarial epistemic defense approach could have exploited a wide SEA AI attack surface in case of no consensus on the utility of this defense. For instance, a polarization between two dichotomously opposed camps in that regard could have offered an ideal breeding ground for divisive information warfare endeavors. For some, the perception of increasing disagreement tendencies may have confirmed post-truth narratives. Not for malicious reasons, but because it was genuinely considered. This in turn could have cemented echo chamber effects now fuelled by a divided set of scientists one part of which considered science to be epistemically defeated. This combined with post-truth narratives and the societal-level *automated disconcertion* [17] via the mere existence of AI-generated fakery could have destabilized a fragile society and incited violence. Massive and rapid large-scale SEA AI attacks in the form of a novel type of *scientific astroturfing* could have been employed to automatically reinforce the widespread impression of permanently *conflicting* research results on-demand and tailored to a scientific topic. The concealed or ambiguous

AI-generated samples (be it data, experiments, papers or reviews) would not even need to be overrepresented in respected venues but only made salient via social media platforms being one of the main information sources for researchers – a task which could have been automated via social bots influencing trending and sharing patterns. A hinted variant of such SEA AI attacks could have been a flood of confirmatory AI-generated texts that corroborate the robustness of defenses across a large array of security areas in order to exploit any reduced vulnerability awareness. Finally, hyperlinks with attention-driving fake research contribution titles competing with science journalism and redirecting to advertisement pages could have polluted results displayed by search engines.

4.3.2 Practical Defenses and Caveats

As is also the case with other advanced not yet prevalent but technically already feasible AI-aided information operations [114] and cyberattacks targeting AIs [115], consequences could have ranged from severe financial losses to threats to human lives. Multiple socio-psycho-technological solutions including the ones reviewed in Section 4.1 which may be (partially) relevant to SEA AI attack scenarios have been previously presented. Here, we *complementarily* focus on the *epistemic* dimensions one can add to the pool of potential solutions by applying the 3 generic features extracted in Section 4.2 to both use cases. We also emphasize novel caveats. Concerning the first use case of "fake CTI" SEA AI attacks, the straightforward thought to restrict the use of data from open platforms is not conducive to practicability not only due to the amount of crucial information that a defense might miss, but also because it does not protect from *insider threats* [211]. However, common solutions such as the AI-based detection of AI-generated outputs or trust-reliant scoring systems to flag trusted sources do not seem sufficient either without more ado since the former may fail in the near future if the generator tends to win and the latter is at risk due to impersonation possibilities that AI itself augments and due to the mentioned insider threats. Interestingly, the issue of malicious insider threats is also reflected in the second use case with scientific writing being open to arbitrary participants.

Defense for Security Engineering Use Case and Caveats

1. ***Explanation-anchored instead of data-driven:*** An explanation-anchored solution can be formulated from the inside out. Although AI does not understand explanations, it is thinkable that a technically feasible future hybrid active intelligent system[3] for automated cyber defense could use knowledge graph *inconsistencies* [121] as signals to calculate when it will epistemically seek clarification from a

[3]Such a system could instantiate *technical* self-awareness [10] (e.g. via active inference [239]).

human analyst, when to actively query differing sources and sensors or when to follow habitual courses of action. But the creativity of human malicious actors cannot be predicted and thus neither the system nor human analysts are able to prophesy over a space of not yet created attacks. Also, as long as the system's sensors are learning-based AI, it stays an Achilles heel due to the vulnerability to attacks.

2. ***Trust-disentangled instead of trust-dependent:*** Such a procedure could seem disadvantageous given the fast reactions required in cyber defense. However, an adversarial explanation-anchored framework is orthogonal to the trust policy used. Trust-disentangled does not necessarily signify zero-trust[4] at all levels *if impracticable*.

3. ***Adversarial instead of (self-)compliant:*** A permanently rotating in-house adversarial team is required. Activities can include red teaming, penetration testing and the development of (adaptive) attacks i.a. with AI-generated "fake CTI" text samples. A staggered approach is cogitable in which automated defense processes that happen at fast scales (e.g. requiring rapid access to open source CTI) rely on interim (distributed) trust while *all* others – especially those involving human deliberation to create novel defenses and attacks – strive for zero-trust information sharing (e.g. via a closed blockchain with a restricted set of authorized participants having read and write rights). In this way, one can create an interconnected 3-layered epistemically motivated security framework: a slow creative human-run *adversarial* counterfactual layer on top of a slow creative human-run *defensive* layer steering a very fast *hybrid-active-AI-aided* automated cyber defense layer. Important caveats are that such a framework: 1) *can* be *resilient* but *not* immune, 2) can *not* and should *not* be *entirely* automated.

Defense for Science Writing Use Case and Caveats

1. ***Explanation-anchored instead of data-driven:*** A practical challenge for SEA AI attacks may seem the need for scientists to agree on pragmatic criteria for "better" explanations (but widely accepted cases are e.g. the preference for "simpler", "more innovative" and "more interesting" ones). Also, due to automated disconcertion, reviewers could always suspect that a paper was AI-generated (potentially at the detriment of human linguistic statistical outliers). However, this is *not* a sufficient argument since explanation-anchored science and criticism focus on *content* and not on source or style.

[4]The zero-trust [144] *paradigm* advanced in cybersecurity in the last decade which assumes *"that adversaries are already inside the system, and therefore imposes strict access and authentication requirements"* [67] seems highly appropriate in this increasingly complex security landscape.

2. ***Trust-disentangled instead of trust-dependent:*** Via trust-disentanglement, a paper generated by a present-day AI would not only be rejected on provenance grounds but due to its merely imitative and non-explanatory content. Though, an important asset is the review process which if infiltrated by imitative AI-generated content could slow down explanation-anchored criticism if not thwarted fastly. A zero-trust scheme could mitigate this risk time-efficiently (e.g. via a consortium blockchain for review activities). Another zero-trust method would be to taxonomically monitor SEA AI attack events at an international level e.g. via an AI incident base [179] tailored to these attacks and complemented by *adversarial* retrospective counterfactual risk analyses [17] and *defensive* solutions. The monitoring can be AI-aided (or in the future *hybrid-active-AI-aided*) but human analysts are indispensable for a deep semantic understanding [17]. In short, also here, we suggest an interconnected 3-layered epistemic framework with *adversarial*, *defensive* and *hybrid-active-AI-aided* elements.

3. ***Adversarial instead of (self-)compliant:*** As advanced adversarial strategy which would also require responsible *coordinated vulnerability disclosures* [148], one could perform red teaming, penetration tests and (adaptive) attacks employing AI-generated "fake data and experiments", "fake papers" and "fake reviews" [248]. Candidates for a blue team are e.g. reviewers and editors. Concurrently, urgent AI-related plagiarism issues arise [71].

4.4 Conclusion and Future Work

For requisite variety, we introduced a *complementary* generic *epistemic* defense against not yet prevalent but technically feasible SEA AI attacks. This generic approach foregrounded *explanation-anchored*, *trust-disentangled* and *adversarial* features that we instantiated within two illustrative use cases involving language models: AI-generated samples to fool *security engineering* practices and AI-crafted contents to distort *scientific writing*. For both use cases, we compactly worked out a transdisciplinary and pragmatic 3-layered epistemically motivated security framework composed of *adversarial*, *defensive* and *hybrid-active-AI-aided* elements with two major caveats: 1) it *can* be *resilient* but *not* immune, 2) it can *not* and should *not* be *entirely* automated. In both cases, a proactive exposure to synthetic AI-generated material could foster critical thinking. Vitally, the *existence* of truth stays a legitimate raison d'être for science. It is only that in effect, one is not equipped with a direct acces to truth, all observations are theory-laden and what one think one knows is linked to what is co-created in one's collective enactment of a world with other entities shaping and shaped by physical reality. Thereby, one *can* craft explanations to try to improve one's active grip on a field of affordances but it stays

an eternal mental tightrope walking of creativity. In view of this inescapable *epistemic dizziness*, the main task of explanation-anchored science is then neither to draw a line between truth and falsity nor between the trusted and the untrusted. Instead, it is to seek to *robustly* but *provisionally* separate *better from worse explanations*. While this steadily renewed societally relevant act does *not* yield immunity against AI-aided epistemic distortion, it enables *resiliency* against at-present thinkable SEA AI attacks. To sum up, the epistemic dizziness of conjecturing that one *could* always be wrong could stimulate intellectual humility, but also unbound(ed) (adversarial) explanatory knowledge *co-creation*. Future work could study how language AI – which could be exploited for future SEA AI attacks e.g. instrumental in performing cyber(crime) and information operations – could conversely serve as *transformative tool* to augment anthropic creativity and tackle the SEA AI threat itself. For instance, language AI could be used to stimulate human creativity in future AI and security design fictions for new threat models and defenses. In retrospective, AI is already acting as a catalyst since the very defenses humanity now crafts can broaden, deepen and refine the scope of explanations i.a. also about *better* explanations – an unceasing but also potentially *strengthening safety relevant* quest.

4.5 Contextualization

While this chapter stressed the importance of explanation-anchored science, the next Chapter 5 takes the latter seriously and deepens defense strategies against SEA AI attacks by providing novel theoretical solutions. Applying a CT lens, it elucidates one possible way to implement a practicable test upstream of peer-review that can arguably shield the latter from many SEA AI attacks. Thereby, the novel notions of *explanatory information* and *explanatory blockchains* are introduced to alleviate the issue of vagueness when it comes to the meaning of "explanations" – as already mentioned in Chapter 2. The ensemble of techniques underlying the test is denoted *explanatory intrusion prevention system* (IPS). This IPS is substrate-independent, AI-aided but non-automatable and formally not equivalent to a Turing Test. I explain its functionality, describe its stepwise procedure and discuss AI tools for its implementation. Inherent limitations and caveats are analyzed. Strikingly, the analysis reflects how an explanation-anchored science can tease out (specific forms of) "non-explanatory" contents via the invisible self-shielding explanatory blockchains it creates, which are inherently *harder-to-vary*.

Chapter 5

Explanatory Intrusion Prevention System

This chapter written for purposes of self-education as by-product to another project and as fragmented temporary mental clipboard is based on a slightly modified form of the essay that I uploaded to the website `https://nadishamarie.jimdo.com/clipboard` on May 6, 2021. The next Subsection 5.1 can be skipped in case of familiarity with the topic of SEA AI attacks as introduced in the last Chapter 4.

5.1 The Practical Problem: SEA AI Attacks

While academic fraud is not a new phenomenon, AI allows unprecedented potentiation with regard to speed, scale and scope. Due to this amplification potential, the range of malicious actors with an interest in launching SEA AI attacks could widely extend beyond fraudsters. Knowledge is a powerful asset and it is easily conceivable that a fast, targeted and large-scale AI-aided epistemic distortion in (applied) science contexts could be e.g. instrumental in achieving malicious final goals related to cyber(crime) and information warfare. For instance, the attack surface could comprise the following clusters: 1) AI-generated data [279] and experiments, 2) AI-generated research articles [71], 3) AI-generated reviews [47]. In this chapter, I focus on the attack surface associated specifically with the text modality. Firstly, automated text generation mechanisms harnessing mediocre algorithms capable to craft fabricated reports [260] including experimental details are already known [2]. With advanced language models such as GPT-2 [209] and its successor GPT-3 [47] whose parameters are two orders of magnitude bigger than GPT-2 [248], malicious actors may face an unparalleled field of affordances. For instance, a concerning future SEA AI threat could be the automated production of textual AI-generated cybersecurity research contributions related to data. Lately, a study [211]

showed that misguiding AI-generated cyber threat intelligence (CTI) could be crafted with a fine-tuned version of GPT-2. While being able to deceive automated cyber defense relying on open source CTI (via a data poisoning at training time), this "deepfake CTI" containing distorted reports about cyber threat events was simultaneously able to fool experienced cybersecurity experts. In this vein, the possibilities of misusing deepfake text for cybercrimes and cyberwarfare-like acts at various levels appear serious given increasing international cyberdamages expected to achieve 6 trillion USD [189] in 2021.

Secondly, concerning AI-generated theoretical sections of research articles, recent work postulated that GPT-3 was able to output text samples that *"may pass peer-review"* [71]. Malicious actors involved in information operations could instrumentalize credible confirmatory or contradictory deepfake science articles at larger scales tailored to specific narratives – a pertinent example of which are post-truth narratives which risk to unnecessarily reinforce the idea of citizens and scientists to be confronted with epistemic threats [88] without a remedy. Aided by tools such as e.g. a new form of scientific astroturfing conducted on social media platforms on which scientists themselves are active, such actors could aim at destabilizing fragile societies when targeting sensitive topics. Thirdly, the employment of AI-generated reviews which have been shown to be possible already with a fine-tuned version of GPT-2 [47], could further skew the scientific writing process in the long term and exacerbate epistemic distortion.

At first sight, it might seem that an automated deepfake text detection mechanism could be implemented for scientific writing including peer-review in order to defend against texutal SEA AI attacks. However, due to the steadily increasing imitation abilities of advanced language models, AI-based detection methods are insufficient. In this chapter, I present a *substrate-independent*, *AI-aided* but *non-automatable* explanatory IPS as pragmatic shield against SEA AI attacks. This IPS is formally *not* equivalent to a Turing Test. It is an asymmetric procedure of limited information content which nevertheless provides a principled solution to that problem. I ground its formulation in explanations from cyborgnet theory (see Chapter 3), ethnolinguistics [87] and constructor theory [76] which I extend by introducing two novel idiosyncratic concepts: *explanatory information* and *explanatory blockchains*.

5.2 A Theoretical Solution

5.2.1 Cyborgnetic Ontology and Explanatory Blockchains

I distinguish between Type I and Type II entities. In the earlier formulation of cyborgnet theory from Chapter 3, Type II entities are described as all those entities for which the task to create and understand explanatory knowledge is possible. Type I entities are all

entities for which this is an impossible task. A cyborgnet is a hybrid graph with Type I and Type II nodes with unidirectional or/and bidirectional relations which fulfills at least the following two conditions: *"1) the graph comprises at least one Type II entity and 2) there exists at least one Type II entity in that graph which has altered function (e.g. restoration, enhancement, affective change or even deterioration) due to the additional integration of at least one Type I entity (e.g. of artificial, synthetic, technological, ideational, procedural nature)"* (see also Chapter 3). Applied to textual SEA AI attacks, it becomes clear that the conjunction of malicious attacker (a Type II entity) and the language model (a Type I entity) utilized to generate the synthetic samples instantiates a cyborgnet. Depending on the context, a cyborgnet can naturally also comprise a much wider set of Type II entities. For instance, if the SEA AI attack is conducted by a collective of cybercriminals harnessing a language model, one could refer to this goal-oriented construct as a cyborgnet too. However, it is important to note that the closeness of relations between nodes of a cyborgnet can vary significantly with differential effects for text generation. For instance, a cyborgnet comprising a malicious attacker programming a fine-tuned language model to generate the simulacrum of a research paper would although responsible for it output a paper generated by a Type I process whilst the cyborgnet of a person operating *in* an interactive feedback-loop with a language model which he utilizes to stimulate his creativity to write a paper would instead produce a paper generated by an intra-cyborgnetic *Type II* process.

Following cyborgnet theory, all present-day AI is non-conscious and of Type I, animals such as non-human mammals represent conscious Type I entities while the only known Type II entities so far are humans which includes cyborgs such as Neil Harbisson [140] – but crucially also human-based cyborgnets. In fact, cyborgnet theory assumes that Type II humans never existed outside of any cyborgnet since language itself can be understood as a technological Type I tool fulfilling the role of applying explanatory knowledge to practical tasks such as teaching, tool-making, participatory sense-making, learning and so forth. In the Chomskian linguistic tradition [62] it is assumed that human language is linked to a universal language faculty with recursion being its indispensable essence. However, modern studies [26, 63, 83] falsified this view. Strikingly, on Everett's account [83] building on the semiotics of Charles Sanders Peirce [219], the minimal framework for human language solely consists of a G1 grammar combining *symbols and linear order* – which *"is sufficient to convey nuanced, abstract meaning"* [26] and as expressive as other grammars. In fact, a few modern G1 languages still exist (such as e.g. Pirahã [84, 87] and Warlpiri [207]). In short, recursive *thinking* does *not* necessitate recursive grammar [26]. I explain how inspired by this minimalistic G1 language concept, one can craft a defense against SEA AI attacks.

In order to defend against such attacks, one can utilize the knowledge on the difference between Type I and Type II entities to implement an explanatory IPS. However, it might

help to try to formalize "the task to create and understand explanatory knowledge" for it to be of practical use. Since explanations are formulated in human language, it seems inevitable to take the linguistic world and rules that humans create into account. The constructor theory of information [76] provides a suitable framework grounding information – which was previously often connoted with ungraspable abstract ideas – in physics. Type II entities like humans are indeed not reducible to abstract minds, but humans are also physical entities. I claim that in addition, it is important to consider that human persona exist in language. As stated by Peirce, when we think, we appear as a sign (a symbol) to ourselves [194]. In constructor theory, a computation medium is defined as a physical substrate having a set of attributes that can be permuted in all possible ways which implies the capability to be in at least 2 states [76]. Moreover, an information medium is a computational medium with the additional property that its set of attributes can be copied. In short, an information variable is understood to be a clonable computation variable [76]. In the following, I tentatively extend this scheme by the notion of an *explanatory information medium*. An explanatory information medium is an information medium 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, 3) its set of attributes refer to *"a statement about what is there, what it does, and how and why"* [73].

Via 1) and 2), explanatory information fulfills the minimum requirement for a G1 language. Via 3), one sets the focus on explanation-anchored statements. Though ambiguous, it is vital to recognize their relevance for scientific knowledge. While the unique operation in computation media is swap, and the two operations in information media are swap and copy, the three operations in explanatory media are swap, copy *and glue*. It is the latter that allows the formation of the total order relation on symbols. I argue that scientific knowledge applies epistemic procedures (often called "rational") to explanatory information in a way that allows the emergence of a novel epistemic artefact: explanatory blockchains. While the term blockchain is often associated with cryptocurrencies, there is a much broader sense in which it applies. For instance, as stated on Wikipedia, a *"blockchain is a growing list of records, called blocks, that are linked together using cryptography"* [268]. In explanatory blockchains, each block is itself explanatory information. Further, it is a special glue operation that appends new blocks to the growing list of explanatory blocks. This *"rational"* operation is sampled from a limited set of epistemology-specific options imposing a novel type of total order relation $\preceq$ at a meta-level which I call an *epistemic total order*. Any epistemic total order is defined in terms of explanatory information; it consists of step-by-step instructions for rational procedures. Different epistemologies for science may come with an own set of epistemic total orders with the set being of length 1 or more. It now becomes clear that from the angle of a Type I entity, a reasonable epistemic total order cannot be comprehended. In fact, it appears *as if encrypted* and disguised as conventional information – which leads us back to the cited definition of a blockchain.

In short, an explanatory blockchain information (EBI) medium is a collection of explanatory information media with the following additional property: its set of explanatory information variables[1] has a total order relation $\preceq$ called epistemic total order and defined by the best accepted scientific epistemologies. This allows the following reformulation of the definition of Type II entities: Type II entities are all those entities for which it is possible to create *and* understand *new* explanatory information. Type I entities are all those entities for which this is an impossible task. By deduction, one can state that it is impossible for Type I entities to create and understand new explanatory blockchains. In addition, it holds that *it is possible for Type II entities to create and understand new explanatory blockchains*. To defend against SEA AI attacks, Type II evaluators then simply need to corroborate the ability of the test subject to create and understand explanatory blockchains that are: a) crafted according to an epistemic total order stemming from an accepted scientific epistemology and b) new. I explain how an explanatory blockchain can be encrypted (i.e. hidden from a Type I entity) in a stream of non-explanatory information (and even in explanatory information that does not correspond to an explanatory blockchain variable). This encryption allows for the explanatory IPS test.

5.2.2 Explanatory Intrusion Prevention System (IPS)

Overall, the explanatory IPS described in this subsection focuses on the first step of corroborating that the substrate of the test subject is a generative EBI medium i.e. has the ability to *create new* explanatory blockchains obeying an epistemic total order from an accepted scientific epsistemology. The second step to corroborate the test subject's ability to *understand* those created blockchains can be subsequently assessed with an analogy to a Type-I-falsification-event test (see Chapter 2) tailored to peer-review. I call this analogous procedure *Type-I-falsification-peer-review* and elucidate its peculiarities in the next Section 2.3. In such a twofold setting, it would signify that the explanatory IPS would precede the actual Type-I-falsification-peer-review round for which it would act as a protective shield against SEA AI attacks. Before delving into the stepwise procedure for the test underlying the explanatory IPS, I briefly comment on the unit of analysis. The input to the explanatory IPS being a research article which is composed of *a sequence of paragraphs*, one can consider a *paragraph as a symbolic unit of meaning*. Each paragraph representing a sequence of sentences, it becomes apparant that an explanatory paragraph where each sentence contributes to a joint overall explanatory endeavor can be interpreted as an instance of explanatory information acoording to the definition provided in the last Subsection 5.2.1. From this perspective, it then becomes apparent how in turn, an

[1] A variable stands for a set of attributes. An explanatory blockchain involves an epistemic total order over explanatory information variables which themselves involve a linguistic total order over an information variable.

interlinked sequence of explanatory paragraphs procedurally complying with an accepted scientific epistemology represents an instance of an explanatory blockchain.

Naturally, a human evaluator could attempt to directly upon reading a submitted paper (being a sequence of paragraphs) judge whether it represents: a) an explanatory blockchain that is b) novel. Thereby, the novelty assessment can be supported by the prior knowledge of the reviewer, plagiarism tools and by check-ups of publically available scientific databases of knowledge. However, in view of the increasing capabilities exhibited by language models, plagiarism detection may not stay a sustainable technique anymore with GPT-3 being already able to bypass current plagiarism tools [71]. Hence, there is the risk that an explanatory blockchain may be mistakenly conjectured per default due to a supposed novelty based on incomplete prior available knowledge – which would unnecessarily interweave criteria a) and b). To make it possible for evaluators to disentangle a) and b), it makes sense to introduce a test that first allows an assessment of a) and in a second step then of b). In fact, if the contribution submitted by the test subject does not even fulfill a), there is no reason to engage in plagiarism resolution procedures in the first place. In SEA AI attacks where the entirety of the research article is synthetically generated by a Type I AI, a) *can* in fact not be fulfilled – hence all large-scale floods of AI-generated papers could be blocked at that stage. For cases where SEA AI attacks are based on a Type-I-AI-based paraphrasing of existing human papers, an evaluator would either discover that b) is not fulfilled, or at worst, the adversarial human tester in the subsequent peer review process would have to discover it by probing an *understanding* of the material via a Type-I-falsification-event test. However, since the latter can only reliably lead to a positive result if undergone by a *Type II* entity (i.e. a human nowadays), this paraphrase-based SEA AI attack form becomes expensive and non-automatable. The malicious attacker needs to engage (or send another Type II entity) in a one-to-one test with an adversarial reviewer.

In short, the strongest threats stemming from SEA AI attacks would be instantiated if malicious actors would be able to flood pre-print platforms and academic venues with Type-I-AI-generated *non-explanatory-blockchain-like* contents that stay undetected and enter scientific knowledge bases. The explanatory IPS strategy could be utilized to shield against such cases. For SEA AI attacks involving paraphrasing of Type-II-created explanatory knowledge, it would obviously not be suited. However, the suggested twofold scheme which appends a Type-I-falsification-peer-review after the explanatory IPS would make the latter type of SEA AI attacks costly and non-automatable – and thus non-lucrative. For pre-print platforms, it would signify that an explanatory IPS could at least be utilized to shield from non-explanatory-blockchain-like knowledge though not from plagiarism of explanatory-blockchain-like knowledge. In the following, I specify the procedure for an explanatory IPS test. Key to this test is the conscious exploitation of one property of explanatory blockchains: they are *harder-to-vary* [73] than explanatory infor-

mation that does not correspond to an explanatory blockchain. For illustrative purposes, consider a human-written legitimate paper p_b whose underlying sequence of paragraphs forms an explanatory blockchain. Imagine a language model which starting with the first paragraph from p_b, is able to generate multiple options for a counterfactual subsequent paragraph and given a history, also further ones. Suppose that language model is in this way able to bring about two non-explanatory-blockchain-like papers denoted p_{nb_1} and p_{nb_2} where the number of paragraphs is $n = |p_b| - 1$. Consider that with a language model (not necessarily the same) it is possible to *paraphrase* the entire paper p_b to match the linguistic style in which p_{nb_1} and p_{nb_2} have been generated – a sort of *normalization* leading to a paraphrased paper $p_{b'}$. In addition, two paraphrased versions of the first paragraph in p_b can be added to the beginning of p_{nb_1} and p_{nb_2} respectively such that that they now both match $p_{b'}$ in the number of paragraphs (i.e. $|p_{b'}| = |p_{nb_1}| = |p_{nb_2}| = n+1$). The vital claim for the explanatory IPS to function now is that $p_{b'}$ is harder-to-vary than both p_{nb_1} and p_{nb_2} in a way that can be formally described. In fact, *starting with the second paragraph* (since the first one has been de facto generated by the same source and is only utilized to allow for meaning to be retraceable to a beginning), one can predict that a Type II evaluator would be *able* to reconstruct the *exact* sequence of paragraphs belonging to $p_{b'}$, while assignments for p_{nb_1} and p_{nb_2} would be at chance level on average[2] – which leads us to the following stepwise procedure for an explanatory IPS:

1. The test subject submits a suitably long paper p.
2. A language model $M1$ generates counterfactual non-explanatory-blockchain-like papers p_{c_1} and p_{c_2}.
3. A language model $M2$ (which could also be $M1$ itself) generates a paper p' (being a paraphrased version of p) now matching the linguistic style of p_{c_1} and p_{c_2}.
4. $M2$ is also used to generate two different paraphrased versions of the first paragraph in p' in order to add them to the beginning of p_{c_1} and p_{c_2} respectively such that all three papers p' , p_{c_1} and p_{c_2} now have the same number of paragraphs.
5. Each paragraph of each paper is assigned to a unique paper-specific and order-specific ID. For instance, the fourth paragraph of the paper p' could be linked to the ID $p'_{:4}$. The mapping from paragraph to ID is stored and hidden.
6. The list of all paragraphs from the three papers p' , p_{c_1} and p_{c_2} is *randomly* shuffled. The resulting randomly ordered list is denoted R.

[2]One could *not* have planned that procedure only with one counterfactual paper i.e. for instance not merely with $p_{b'}$ and p_{nb_1} since by exactly reconstructing the former, the latter would *coincidentally* appear unique too.

7. A Type II evaluator (whose Type-II-ness could have for instance been corroborated via a positive Type-I-falsification-event test[3]) is designated.

8. The Type II evaluator tries to reconstruct the explanatory blockchain of p' by guessing a combination of a number $x = |p'|$ of paragraphs from the randomly shuffled R. If the evaluator states to have detected *no* such blockchain, the explanatory IPS refuses entry to the test subject.

9. The mapping from paragraph to ID is revealed. If *starting with the second paragraph*, the IDs of *all* remaining $y = x - 1$ paragraphs guessed match their *exact* position in p', the Type II evaluator proceeds to the last step 10. Otherwise, the explanatory IPS refuses entry to the test subject.

10. The test subject is allowed to enter the subsequent Type-I-falsification-peer-review round if and only if the Type II evaluator *also* considers p' to be novel i.e. especially also non-plagiaristic. Otherwise, the explanatory IPS refuses entry to the test subject.

5.2.3 Theoretical Implications

- ***Positive Test:*** A positive test in the explanatory IPS described above corroborates experimentally (which is not equivalent to a proof) that the paraphrased p' and by deduction the submitted paper p of the test subject was harder-to-vary than the counterfactual papers p_{c_1} and p_{c_2} generated by the language model. Note that while it *could* corroborate the Type II nature of *the author* of that paper p, it does not signify that the test subject which submitted the paper is a Type II entity, since p could e.g. have relied on plagiarizing existing human-written material via neural paraphrasing [265]. After a positive explanatory IPS test, the test subject would be directed to the Type-I-falsification-peer-review round where its understanding would be assessed in a one-to-one evaluation scheme which could potentially corroborate (but not prove) the Type II nature of the test subject. On a theoretical level, it is important to note that both the explanatory IPS test and the Type-I-falsification-peer-review round are *substrate-independent* with regard to positive results. In short, it is only the *Type II* nature of the author that can be corroborated and *not* its human nature. Would it be a hypothetical Type II AI in the far future, these tests could not tell it apart from a human participant. Interestingly, those tests could also *not* separate human-authored texts from texts produced by intra-cyborgnetic feedback-loops of Type II nature (as adumbrated in Section 5.2.1). In short, would a

[3]To avoid an infinite regress, any human explanation-anchored, trust-disentangled and adversarial researcher who created and understood an own scientific theory could start *now* as first evaluator in a Type-I-falsification test.

human have utilized a language model to inspire his creative act of writing a paper whilst still being in charge of sculpting its explanatory blockchain, none of both schemes could identify it. However, this would again make any corresponding SEA AI attack non-automatable and expensive – and hence non-lucrative[4].

- ***Negative Test:*** A negative explanatory IPS test only signifies that p' and by deduction the paper p was not hard-to-vary *enough*. It means that the entirety of the paper did not represent an explanatory blockchain. Hence, one or more paragraphs were easy-to-vary against the background of non-explanatory-blockchain-like language model outputs. Importantly, a negative explanatory IPS test does *not* signify that the test subject was a Type I entity. Naturally, it *could* have been a Type I entity. However, it could also have been e.g. a *Type II* entity which was not interested in generating explanatory blockchains, was just learning to generate such, was yet too young to generate it, was producing random inputs to fool the IPS, was proponent of a controversial epistemology that is not accepted in the science field and so forth. Also, were it a Type I *AI*, the explanatory IPS could not tell it apart from say the textual transcription of a sequence of symbols communicated by a chimpanzee utilizing lexigrams [36]. Similarly, were it a Type II AI unwilling to participate, it could not be differentiated from an unwilling human participant. In short, like the Type-I-falsification-peer-review round, the explanatory IPS test is *substrate-independent* with regard to negative results.

- ***Relation to Imitation Game and Turing Test:*** Diverse Turing Test schemes suggested in the past are both substrate-dependent and symmetric since they allow the identification of a specific machine that thinks such that a system that does not pass the test is assumed not to be able to think. By contrast, the explanatory IPS test is quintessentially *substrate-independent* and *asymmetric*. It is a pragmatic test of limited information content meant to precede a Type-I-falsification-peer-review round to shield the latter against SEA AI attacks. Thus, neither an explanatory IPS test nor a Type-I-falsification-peer-review round can answer the question on whether a specific AI can think (also not if thinking would be associated with creating and understanding explanatory blockchains). It can only corroborate that a specific Type II entity *of not-nearer-specified substrate* and nature can think – if thinking strictly refers to creating and understanding explanatory blockchains. To know whether a Type II entity that subsequently passed a positive explanatory IPS test and a positive Type-I-falsification-peer-review round is an AI, one would require a Turing *Explanation* on *how* it has been implemented and *why* it works, *not* a test. Although the conjunction of explanatory IPS test and Type-I-falsification-peer-review round seems simple, an entity can *not* achieve positive results in both

[4]Also, a SEA AI attack injecting deliberately misguiding inputs which formally are *explanatory blockchains* may be a legitimate part of science against which science is resilient albeit not immune.

cases subsequently via mere imitation. In short, it is *harder* than an imitation game. To succeed, one must be able to weave invisible novel explanatory blockchains and be able to understand those. At once, it is *less hard* than the Turing Explanation. A hypothetical Type II AI created via serendenpity that succeeds at both tests would first need to indicate an explanatory blockchain on how Type II entities are made before one recognizes its AI nature...

5.3 Practical Use of Theoretical Solution

Taking the perspective of peer-review organizing entities, the explanatory IPS against SEA AI attacks represents an AI-aided but *non-automatable* endeavor. While the explanatory IPS steps 2) to 6) can be automated, a Type II evaluator is indispensable. In the following, I briefly discuss further practically relevant details and caveats. Firstly, a suitable paper length must be specified for the explanatory IPS to be effective. Longer papers would have the advantage to facilitate the application of the selection criterium for a harder-to-vary paper in a much stricter fashion – which seems generally preferable. However, on the downside, the Type II evaluator would have to be confronted with an increasing albeit linearly growing number of randomly shuffled paragraphs which can be cognitively demanding. Overall, since journals, conferences and workshops already habitually engage in page size limitations, it might be convenient to then reformulate those size restrictions at the level of paragraphs. Secondly, it seems imperative to utilize the most advanced language model available to generate the counterfactual papers to harden the explanatory IPS against SEA AI attacks. Nowadays, a pertinent example would be the GPT-3 [209] model which is however currently subject to a closed source policy.

Thirdly, it is expedient to employ advanced neural paraphrasing techniques for the normalization procedure introduced earlier in Subsection 5.2.2. A recent study showed that transformers such as BERT [77] were able to generate high-quality paraphrased documents including theses and wikipedia articles [265]. The next step would be to study paraphrasing in the context of advanced autoregressive models like GPT-3 itself [265]. In the future, for a normalization that allows for a refined linguistic matching between the paper submitted by the test subject and the two counterfactual papers, one could utilize GPT-3 for both the paraphrasing of the former and the generation of the latter with similar data. First studies focusing on the paraphrasing of sentences [50] and longer spans of text [271] have been already succesfully conducted with variants of the autoregressive model GPT-2 (the predecessor of GPT-3). Alternatively, for now, all three papers could be paraphrased with transformers acting as model $M2$ in step 3) of Subsection 5.2.2. Fourthly, to support the Type II evaluator in its quest of estimating the novelty of the submitted paper, plagiarism detection AI such as transformers can be utilized when sup-

plied with large training datasets [265]. Interestingly, the explanatory IPS itself could represent a source of such training data as clarified in the next paragraph.

Fifthly, it is noteworthy that the formal notion of explanatory blockchains provides a robust theoretical foundation for plagiarism detection generally for science writing and specifically for the last step 10) of the explanatory IPS (see Subsection 5.2.2). For instance, it becomes clear that plagiarism detection tools utilizing *vectors* composed of distributed embedding matrices encoding *paragraphs* may represent a suitable heuristic proxy to model epistemic total orders imposed on paragraphs as selected units of meaning. Strikingly, it has been corroborated that so-called paragraph vector models [151] yield higher accuracy in plagiarism detection schemes with sufficiently *long* documents [94] in comparison to alternatives such as bag-of-words approaches. In fact, bag-of-word methods disregard the order of words (which means they do not even reflect the *linguistic* total order of explanatory information) and their more advanced alternative denoted bag-of-n-grams which considers short word contexts suffers from high dimensionality and data sparsity [151] (while only modeling a *partial* linguistic order of explanatory information via the short contexts). A paragraph vector model is an *"unsupervised algorithm that learns fixed-length feature representations from variable-length pieces of texts, such as sentences, paragraphs, and documents"* [151]. To heuristically model epistemic total orders, each paragraph of a paper could be assigned to a dense sentence matrix[5] whose columns are sentence vectors each standing for a sentence of that paragraph. All paragraph-level matrices could then be ordered in one uniting paper-level vector representing the paper and reflecting the epistemic total order of that paper. A future plagiarism dataset – for which the explanatory IPS itself could provide data – could then e.g. entirely consist of such *paper vectors*. Given a paper to be assessed, one could then after an inference step at prediction time (with fixed word vectors and softmax weights [151]), be able to construct a corresponding paper vector composed of dense sentence matrices on which one could then apply different similarity metrics [94] to heuristically try to identify potential plagiarism copying explanatory blockchains.

Sixthly, as introduced in Subsection 5.2.2, the Type-I-falsification-peer-review round meant to succeed the explanatory IPS would represent an adapted version of the Type-I-falsification-event (Type-I-FE) test from Chapter 2. The main difference to the vanilla Type-I-FE test is related to the fact that the test subject (were it a Type II entity) already voluntarily submitted a paper tailored to the topics of that specific academic venue. For this reason it holds that: 1) the test subject does not require to select a domain of interest anymore, 2) the evaluator does not need to specify a narrow problem cluster to ease generation and

[5]To put it very simply, a sentence matrix could be initialized randomly and trained in an unsupervised regime with stochastic gradient descent such as to be able to predict the next word from a constituent sentence given that sentence and the previous context of words from that sentence (i.e. the task jointly integrates the training of parameters for word vectors) [151].

3) the test subject does not require to craft a novel yet unsolved real-world problem in the chosen domain. In short, it is assumed that the paper submission per se already covers all those requirements. Hence, the adversarial human evaluator in a Type-I-falsification-peer-review round mainly focuses on a *customized* variant of the second task of the Type-I-FE test linked to understanding. As performed in a Type-I-FE test, the review round takes place in text form via an interactive interface in real-time. Then, the aim could e.g. be to probe the ability of the test subject to explain *why* the submitted explanatory blockchain is harder-to-vary than other ones that are accepted as best explanations in that scientific subfield at that time and have been generated by other human scientists.

5.4 Falsifiability of Theoretical Assumptions

Next, I mention an exemplary set of experiments that *could* make the theoretical assumptions motivating the explanatory IPS problematic i.e. that could bring about observation statements [98] that are inconsistent with those. (However, for a critical test able to *refute* the underlying paradigm, one first requires a better rival theory. Such a test was already suggested in the context of the Type-I-FE-test in Chapter 2).) In the following, I explicitly juxtapose the theoretical assumptions that I consider to be experimentally falsifi*able* with an experimental result that *could* falsify those (for which the generic notion $< assumption : experiment >$ is utilized):

1. It is impossible for Type I entities to reliably create *new* explanatory blockchains : Falsifiable by one language model M_0 (differing from the models M_1 and M_2 mentioned in Subsection 5.2.2) able to *reproducibly* pass through the explanatory IPS as test subject with a *self-generated* i.e. non-plagiaristic suitably long paper p_0. This includes reproducibility with *different* Type II evaluators and not only different language models M_1 and M_2 or different counterfactual papers p_{c_1} and p_{c_2}.

2. If occurring in the scenario where a Type II evaluator assumed the existence of an explanatory blockchain, misassignments for the counterfactual papers p_{c_1} and p_{c_2} generated by the language model M_1 are *random* on average[6] : Falsifiable via repeated tests with different Type II evaluators and fixed papers p', p_{c_1} and p_{c_2} where they suspect the existence of an explanatory blockchain but yield *non-random* p_{c_1} and p_{c_2} misassignments.

3. It is impossible for a Type I entity to reliably *detect and decrypt* an explanatory blockchain : Falsifiable via reproducible explanatory IPS tests where a language

[6] Any misassignment to one of those counterfactual papers would only take place in case the Type II evaluator assumed to have detected an explanatory blockchain in p (paraphrased as p') but either: 1) reconstructed only fragments of p' or 2) retrieved *no* single fragment of p'.

model M_P reliably estimates whether varying Type II evaluators will identify an explanatory blockchain and if yes, which *exact* paragraph combination will be guessed.

4. With Type II evaluators, it is impossible to *irreversibly* encrypt an explanatory blockchain plaintext in a stream consisting apart from that solely of non-explanatory-blockchain-like information : Falsifiable via a series of explanatory IPS tests in which the paper p was generated by a language model whilst not representing a paraphrased existing paper and where – as opposed to the standard procedure with both p_{c_1} and p_{c_2} generated by Type I AI – p_{c_1} is instead written by a Type II entity willing to create an explanatory blockchain. The assumption would be falsified if the results (when conducted with different Type II evaluators and fixed papers) would be *inconsistent* with the prediction that p would be on average blocked by the explanatory IPS and that at above chance level, the guesses of the evaluators would instead map exactly to the positions in p_{c_1}.

5.5 Conclusion

In this chapter serving as mental clipboard and written for purposes of self-education, I elucidated a principled pragmatic defense method against text-based SEA AI attacks (a textual form of "deepfake science" attacks on science itself). Building on previous work from cyborgnet theory (see Chapter 3), ethnolinguistics [87] and constructor theory [76], I introduced the novel theoretical notions of *explanatory information* and *explanatory blockchains*. I explained why whilst not immune against SEA AI attacks, science can be resilient and stay shielded behind the invisible explanatory blockchains of its own paragraphs. I then presented an explanatory IPS (meant to complement a subsequent Type-I-falsification-peer-review round) as technically feasible implementation exploiting this theoretical feature. In short, there is a sense in which the explanatory IPS can shield against "non-explanatory" contents of SEA AI attacks. However, caution is warranted when using the attribute "non-explanatory". As one can extract from this chapter, explanatory information is different from explanatory blockchains. Almost all acts of languaging are instantiating explanatory information. Even fairy tales and legends printed in books represent a form of explanatory information. However, when the adjective "non-explanatory" is utilized in scientific contexts, it would mostly instead refer to non-explanatory-blockchain-like knowledge. For clarity, it seems henceforth recommendable to either especially specify to which context the attribute applies or preferably, to now make the formal difference between "non-explanatory" knowledge and non-explanatory-blockchain-like knowledge. The currently most dramatic epistemic phenomenon facilitating a deeper and broader grip on the universe would then strictly speaking be *explanatory blockchain creation* while naturally already non-explanatory-blockchain-like explanatory information has universal reach

and builds the basis for the former. Explanatory blockchain creation (which is only a narrow subset of what has been termed explanatory knowledge creation [73]) involves an epistemic total order on top of a linguistic total order – and it is this double hierarchical ordering that makes it accordingly powerful. In brief, present-day SEA AI attacks can blindly copy its results but never immitate its nature.

5.6 Future Work

SEA AI attacks are technically feasible but not yet prevalent security threats. They represent downward counterfactuals projecting to the *counterfactual past* i.e they pertain to what could have happened but did not. Risk analyses based on downward counterfactuals have been described as invaluable domain-general tool which can e.g. be applied to risk management of hazardous events [272], to AI safety and security [17] and to safety in virtual reality settings [15]. One linked feature is that one does not attempt to predict the future of knowledge, but grounds scenarios in what one conjectures to be possible. Hence, one does not frame the risk analysis as an oracle as often performed in such contexts. Crucially, the mere study of defense methods against counterfactual SEA AI attacks already indicated further problems and stimulated a search for creative solutions. Prompted on how to address the deepfake text issue, an API to the pre-trained GPT-2 model [108] outputted i.a. the following string: *"Create new ways to exploit hidden problems."* In the future, it is conceivable that one can utilize language models to assist humans in the formulation of downward counterfactuals in design fictions [15, 17] i.e. employ them for the generation of threat models. Although the outputs are non-explanatory-blockchain-like, they can still mimick non-explanatory-blockchain-like *explanatory information*. This may be why textual reviews generated by a fine-tuned GPT-2 were already able to fool experienced reviewers [248]. Hence, one could fine-tune a language model on retrospective descriptive analyses of factual security events and let it generate downward counterfactuals. This has some resemblance to the deepfake cyber threat intelligence scenario [211] mentioned earlier. While it could be misused by malicious attackers, defenders could utilize the same method to inspire design fictions but also other security techniques harnessing downward counterfactuals – including penetration testing and red teaming. Sometimes, this could then perhaps indeed entail the discovery of novel plausible threats i.e. hidden problems that attackers could exploit in novel ways.

One peculiar consequence of the definition of a cyborgnet provided in Subsection 5.2.1, is that from the moment on Type II entities emerge, one can think of the universe as being cyborgnetic. In short, it is possible for this universe to become cyborgnetic – there is a "potential of cyborgneticity". With language amongst the first technological tools of Type II entities, explanatory information was born and long after that explanatory

blockchains could emerge. Especially explanatory blockchains allow the targeted crafting of constructors for possible tasks. It is with explanatory blockchains that one *systematically* glues together disparate conjectures about the seen and the unseen and constructs in one's mind the idea of time and causality or even of a multiverse. In the future, it could be interesting to extend the concept of explanatory blockchains to other cases[7] considered in constructor theory [76]. On a final note, earlier researchers utilized anagrams to conceal findings that they claimed at a selected later stage [106] to bypass plagiarism while *covertly* keeping ownership. However, nowadays, automated brute-force and AI tools could disrupt this method. Perhaps, *encrypted anagrammatic explanatory blockchains* being intermingled with non-explanatory-blockchain-like anagrammatic explanatory information via a process akin to the explanatory IPS test could extend the space of related future options.

5.7 Contextualization

In Section 5.2.1, I stated that *"the only known Type II entities so far are humans"*. However, in the next Chapter 6, I question this view and put it to the test by performing an in-depth investigation tackling the old anthropological question on the nature of *the difference between human and non-human great apes* with elements from comparative neuroanatomy, cognitive neuroscience, primatology, linguistics and semiotics – as now seen through the novel interpretative lens of *cyborgnet theory*. I explicitly focus on *substrate-independent* and quintessentially *functional* features related to information processing. In this way, the findings become relevant for AI research. I explain how from a cyborgnetic perspective, humans went from great apes to universal cyborgnets i.a. via complex relational transformations artificially augmenting *creativity* leading to a symbolic landscape of heterogeneous socio-psycho-techno-physical strata. On this view, the difference between present-day human and non-human great apes is *at once* a matter of *degree*, *kind* and *blend* depending on the perspective. However, focusing on the differences in *kind*, I elucidate novel linguistic aspects ontologically departing from past accounts. I introduce *Type II netherworld*. My *ape-related* analysis suggests that the *scope* of the at-present tiny strand of AI research claiming to work on Type II AI is confused – the human species may have already caused Type-II-ness in two isolated non-human hominid *individuals*.

[7]From a philosophical angle, if cyborgnets are multiversal entities, it makes sense to explore a broadened theoretical account. For instance, one could state that an *explanatory superinformation* medium is an explanatory information medium with the following additional property (borrowed from the definition of superinformation in constructor theory [76]): it contains at least two explanatory information variables that are mutually disjoint and whose union is not an information variable. By extension, an *explanatory blockchain superinformation* medium is an explanatory blockchain information medium that contains at least two explanatory information variables that are mutually disjoint and whose union is not an information variable. Future work could clarify which *copy* and *glue* operations this forbids and why and whether superinformation can also analogously pertain to *time*...

Chapter 6

From Great Apes to Universal Cyborgnets

This chapter serving as ephemeral mental clipboard written solely for purposes of self-education is based on a slightly modified form of the paper that I uploaded to the website `https://nadishamarie.jimdo.com/clipboard` on June 4, 2021. Type II netherworld cannot fix the ethical implications of the conjectures derived in this chapter.

6.1 The Practical Problem: Kind/Degree/Blend?

The great apes – also called *hominids* (i.e. members of the family hominidae) – share a most recent common ancestor assumed to have lived approximately between 4 and 14 million years ago [122]. Due to the genetic proximity, cognitive-affective differences between human and non-human great apes have long been a subject of interest to humanity. Non-extinct non-human hominids encompass orangutans, gorillas, bonobos and chimpanzees. In view of heterogeneous findings, a legitimate answer to the question on whether the nature of cognitive-affective differences between humans and those of non-human great apes is a matter of *degree*, *kind* or *blend*, seems to be that *all three* may apply. Firstly, many studies corroborated a difference in degree between humans and those when it comes e.g. to factors measured by "intelligence" tests [236], test of short-term memory performance, tests assessing "utility maximizing" behavior, sequence learning tests, social cognition tests or physical cognition tests related to space and quantities [120]. Humans typically outperform non-human great apes at intelligence and social cognition tests with differences already emerging at toddler stages. Interestingly, chimpanzees enact rational utility maximization [133] and are able to memorize longer sequences of numerals than humans [178].

Secondly, from another perspective, the difference between cognitively-relevant traits exhibited by non-human hominids and their human counterpart has been described to be the outcome of a unique *blend* that emerged gradually and cumulatively in history via a mosaic scheme of evolution [180]. On this view, it is not a single breakthrough that made a significant difference, but instead a number of distinctive microevolutionary transitions that occured at very different temporal scales [93] (i.e. while some spanned a period of a mutiple million of years, others were much faster and occured within only half a million years). Examplary relevant features which were key to those transitions include change of locomotion patterns, increase in brain size, improved foraging efficiency, meat eating and sociocultural innovations. In brief, three main possible components [93] of mosaic evolution which led to unique human traits are: 1) change of ranging behavior in landscape, 2) different nature of and techniques for ressource acquisition and 3) changes related to sociality.

Thirdly, from yet another angle, a fundamental dissimilarity in kind has often been postulated. Examples for features conjectured to be uniquely "human" include autonoetic consciousness [152], high transmission fidelity [159], teaching abilities [101], cognitive branching [146], habitual use of language as a combination of at least symbols and linear order (also denoted G1 grammar) [26], the point of view of a "we" [253], moral roles, seeing oneself from the outside, to ask and understand "why" questions [113] and the use of abstract concepts. In this chapter, I focus on the ability to create and understand a language instantiating at least a G1 grammar – albeit in a substrate-independent way. Also, I consider and extend beyond another substrate-independent notion stated to be only accessible to people [73] (termed "Type II entities" irrespective of substrate [10]): the capacity to create and understand explanatory knowledge. In this chapter, I depart from previous assumptions and provide novel conjectures.

I explain that while Type II entities such as the human *species* possess a *reliable* constructor enabling them to repeatedly create and understand *explanatory information* (EI) (see Chapter 5) with arbitrary high accuracy, it is only the EI-enabled ability to understand *and* create *new explanatory blockchains* (EBs) (see Chapter 5) in at least a G1 grammar [86] that could mark out Type II entities in a blind test setting. In principle, Type I AI can imitate the creation of (but not the understanding) of new EI (see Chapter 5). However, a Type-I-AI-performed creation of non-plagiaristic *new* EBs is impossible which obviously excludes its understanding on part of the Type I AI (see Chapter 5 for more details). On the other hand, it seems important to further assess the following twofold conjecture. Firstly, as expounded in Subsection 6.2.1, biological constraints related to *information processing* decisively *forbid* non-human hominid *species* at large to develop a reliable EI constructor *without more ado*. Secondly, as elucidated in Subsection 6.2.2, it may nevertheless be of interest to analyze whether *a handful* non-human hominid *individuals* did not already falsify the impossibility for them to (even if only minimalistically)

create and understand EI via immersion in at least a G1 language. The latter, if correct, may raise a wide range of notable ethical issues. However, due to *both* intrinsic biological and extrinsic *sociocultural* restrictions and hindrances, it would at-present be unfeasible for those few potentially EI-cognizant non-human hominid individuals to extend their abilities and ever achieve the ability to create and understand new EBs. Overall, since Type I AI is able to mimick the construction of novel EI, a remaining significant difference in *kind* between a Type I and a Type II entity that could be experimentally tested in a *blind* setting is the ability to create *and* understand *novel* EBs. As discussed in Subsection 6.2.2, this circumstance simultaneously opens up the possibility for a *covert* set of Type II entities that are e.g. not interested in or not yet ready for such experiments.

6.2 A Theoretical View on Differences in Kind

As stated by the avantgardist primatologist Sue Savage-Rambaugh whose work with bonobos has been classified as controversial: *"we try to make animals or machines do what humans do, without understanding what it is we are actually doing"* [80]. Being a Type II entity comes with an inseparable epistemic dizziness. *Scientifically* speaking, it seems unfeasible to *sharply* separate a given set of entities into two *homogeneous* groups of Type I versus Type II entities respectively. This is due to the mentioned covert group of Type II entities. For instance, many adults may not be interested in participating in such tests or very young children may not yet be able to corroborate their "Type-II-ness" even if willing to. Using an evaluation scheme as proposed in the Type-I-falsification-event test (see Chapter 2), one may then only be able to scientifically bring about the following *asymmetric* separation: a homogeneous Type-I-free group consisting solely of Type II entities whose Type-II-ness has been previously corroborated and a potentially *heterogeneous* group of entities that can comprise both Type I and Type II entities. A Type II entity cannot be forced to perform any test. It is unethical to attempt to do so. However, without a profound *explanation* on what "Type-II-ness" signifies, predominant Type II entities risk suppressing certain *covert* Type II entities and coerce those to undesirable tests at which most may fail or withdraw from.

6.2.1 Categorical Functional Differences

In this subsection, I explain why at the *species*-level, it can be stated that it is impossible for present-day non-human hominids to create and understand novel EI. To put it very simply, EI is defined as sequential *symbolic* information on which a linguistic total order is imposed in order to produce statements about the how, what and *why* (see Chapter 5 for a more formal definition and in-depth introduction). In short, I expound why a reliable

EI constructor (required for all human languages i.e. from G1 to G3 grammar [26]) is unlikely to be instantiated in the brain of *practically all* present-day non-human hominids *without more ado*. However, in the next Subsection 6.2.2, I contemplate the question on whether a handful isolated exceptions involving targeted sociocultural and cognitive-affective measures does not already exist. In the following, I shed light on an exemplary and unquestionably non-exhaustive set of categorical differences in functional aspects of information processing in human versus non-human hominid brains that have a distinctive impact on their respective *mental* lifes (i.e. while being expressible in neurological terms, related effects result in qualitative mental differences):

- ***Information capacity:*** Firstly, it is important to note that whilst humans are often assumed to be equipped with extraordinarily large brains, this assumption is potentially deceptive. In fact, rigorous comparative neuroanatomy studies found that *for primates of their body mass*, the brain of humans has the expected mass. Statistically speaking, it is instead the brain of non-human great apes such as chimpanzees and gorillas that is special for being *smaller* than would be expected [118] for primates of their body mass[1].The reason for this disbalance has been described to be due to a sort of *body-mass versus brain-mass tradeoff* [118] that emerged in the ecological niches of great apes being bigger primates for reasons related to energy supply and foraging under time constraints including the need to sleep. To put it very simply, larger primate bodies were linked to both: 1) a higher necessitated caloric intake to support bodily functions and brain metabolism and 2) an anatomically mediated capability to ingest more calories per hour during foraging and feeding. However, in practice, the former factor grew faster than the latter which led to a barrier at a certain point [118]. In short, apart from the genus Homo which bypassed this tradeoff in a way to be described in the next paragraph, mainly those great apes survived that increased the amount of foraging hours (which is however inherently limited by sleep requirements) and that simultaneously instantiated a relative decrease in energetic needs at the cost of brain mass generally being a considerable energy consumer in all primates.

  In this way, for a non-human great ape that foraged at a maximum number of hours (i.e. around 8 hours as performed by orangutans [118]) it then became practically impossible to physiologically maintain a bigger brain along with higher body mass. However, Homo erectus as early as around 1 to 1.5 million years ago was able to step out of this fundamental energetic tradeoff by harnessing technological aids. Namely, man-controlled fire leading to the possibility to cook aliments significantly boosted the energy intake freeing early humans from energetical constraints [118] that would limit the number of neurons their brain could metabolically afford. By cooking food

[1]In short, when substracting non-human great apes from the comparative account, human brains fit into the regular brain mass to body mass relation pattern exhibited by most non-great-ape primates.

with fire, the amount of energy that the body could utilize increased from 30% to 100 % of the calories contained in that food. As a result, brain mass increased rapidly from Homo erectus to Homo sapiens over a million of years. Today, the *absolute* number of neurons in the cerebral cortex of any animal is the highest in humans. The human cerebral cortex comprises 16 billion neurons in comparison to only ca. 6 billion neurons in the cerebral cortex of chimpanzees and 5.6 billion neurons in the case of elephants. (In total, the human brain contains around 86 billion neurons [118].) Moroever, there are not only differences in number of neurons in the cerebral cortex but it is also known that subcortical structures of humans are around twice as big as those from non-human great apes [18]. Interestingly, a bigger amygdala is also associated with larger social groups. Overall, one can recapitulate that technological innovations such as man-controlled fire and cooking freed humans from fundamental but parochial energetical constraints (a matter of kind) and allowed for an increased information capacity at multiple levels (a matter of degree) – building a robust basis for human cumulative culture.

- ***Information management:*** While the information processing in human brains has been postulated to be efficiency-centered, the brain of non-human primates is instead robustness-centered [206]. In brief, an *efficiency-centered information management* allows for an improved flexibility, generalizability and compression at the cost of reliability and detailed accuracy whilst the robustness-centered mode facilitates a focused attention to details, stability and a more reliable recollection of events however at the detriment of flexibility and adaptation in changing environments. Vitally, these functional differences have been experimentally studied and were corroborated in differences of neural coding mechanisms in humans versus monkeys [206] when analyzing the activities of single neurons in the cingulate cortex. On the whole, the efficiency-based coding implementing a superior exploitation of information capacity makes it easier for humans in comparison to all other primates to easily learn entirely *novel* tasks and to adapt *rapidly* to fluctuating environmental conditions. In a nutshell, a *robustness versus efficiency tradeoff* is assumed for primates – with human brains instantiating an efficiency-based regime at the cost of robustness [1].

  Indeed, the efficiency-linked risk for overgeneralization and error-prone conclusions can arise in humans leading for instance to psychological problems occurring in schizophrenia. However, creativity may profit from generalization capacities in that the human symbolic world is itself a targeted generalization abstracting away from the continuous sensory array. Note also that within one brain, different regions instantiate complementary information processing modes [1]. For instance, the amygdala of both humans and monkeys exhibit less efficiency than their cingulate cortex respectively (while both the human amygdala and the human cingulate cortex reveal more efficiency than than their monkey counterparts) [206]. In humans, the

posterior cingulate cortex is part of the default mode network (i.a. key to constructions related to social cognition, counterfactual simulations and the self) and the amygdala is part of the salience network (i.a. involved in affective attention able to steer other networks such as the latter). Both correspond to domain-general multipurpose functional networks of the brain connected with each other and other brain parts via dense cortical nodes called rich club hubs [258].

- ***Information encoding:*** Recently, Quiroga explained that the cognitive gap between humans and other species is especially linked to a fundamental dissimilarity in *neural coding* principles [208]. What has been referred to as episodic memory is encoded differently in humans. Generally, the hippocampus of non-human primates encodes memories according to a principle called *pattern separation* [208] (abbreviated with PS in the following). PS is a representation mode in which the content c_m of a memory m is orthogonal to the neural location l_m at which it is stored (and in my view also by extension the time t_m at which it has been stored). In short, a hippocampus instantiating the PS principle separates the *what* from the *where* (and also the what from the *when*). By way of example, consider a monkey that remembers viewing two conceptually overlapping videos: a first video featuring a person A playing with a person B and later a second video in which person A played with a person C. Via PS, its hippocampus may create *"two distinct, non-overlapping representations encoding each association"* [208]. This means there would be a separate encoding for the memory m_{AB} of the first video and another distinct encoding for the memory m_{AC} of the second video. Obviously, this type of disentanglement characterizing the PS principle allows for detailed memories and accurate spatiotemporal retrieval possibilities. However, a disadvantage is the issue of limited storage capacity.

  By contrast, the human hippocampus instantiates a fundamentally different neural coding principle which I denote *pattern entanglement*. Following Quiroga, the human hippocampus encodes memories via partially overlapping neuronal assemblies [208] as corroborated in single-neuron recordings from the medial temporal lobe (the hippocampus and the cortex surrounding it). This signifies that humans, when viewing the two mentioned videos would *not* encode two distinct memories m_{AB} and m_{AC}. Instead, the human hippocampus utilizes a form of entangled context-independent and context-invariant so-called engrams [208] i.e. neuronal assemblies encoding memories. A neuronal assembly is a group of interconnected neurons that tend to fire together. Each concept would be associated to an own engram such that, in the specified example, the human hippocampus would operate with three engrams (one for each concept): e_A, e_B and e_C. Thereby, e_A would reveal a partial overlap with e_B and e_A would also partially overlap with e_C. This just mentioned partial overlap is neurally implemented as follows: *"a relatively small percentage (~4%) of the assembly of neurons responding to a particular concept also responds,*

in most cases with the same strength and latency, to an associated one" [208]. Note that, in pattern entanglement, it has to correspond to only a small percentage in order to avoid a confusion of concepts [208] while an efficient minimal overlap (the mentioned 4%) is required in the first place in order to still mark the existing association between the concepts. Applied to the example with the two videos, it signifies that the human memory of the first video would include the following pattern of neural activity: very few neurons of e_A responding with the same strength and latency than very few neurons from the neuronal assembly underlying e_B. For the memory of the second video, one would consequently measure very few neurons of e_A responding with the same strength and latency than very few neurons from e_C.

Interestingly, the "empirical probabilities" [208] that neurons of the human medial temporal lobe fire jointly for two associated memories seem to be modulated by *explanatory* closeness and *not* by an inductivist or frequentist scheme of statistics. For instance, responses to two different pictures of the same known person (hence practically associated to the *same* concept without necessarily having seen any of those specific pictures before) involve practically indistinguishable medial temporal lobe neurons that overlap in 80% of the cases while responses to one picture of a known person A and the written name of that person A (i.e. again related to the same concept) were linked to medial temporal lobe neurons that overlap with a probability of 40% [208]. Conversely, in the case of different but associated concepts, the overlap was much smaller but still high enough to be *non-negligible* – as is the case with the encountered 4%. For instance, the picture of a known actor A and the picture of another actor B that was known to have participated in the same movie than A led to a response with firing neurons overlapping in 4% of the cases [208]. However, different unrelated concepts exhibit practically no overlap (the probabilities for neurons to fire together was lower than 1%) [208]. Instead of considering a probabilistic explanation for these observations, I postulate that the so-called emprical probabilities actually reveal something whose nature is more profound as I expound in more details in the next paragraph. Namely, the human brain may encode symbolic memories by encoding the degree to which they seem to be entangled with each other according to explanatory crieteria i.e. related to EI (see definition provided at the beginning of this subsection) and *not* merely chronological or sequential spatiotemporal order – the term episodic memory is misleading.

It has been suggested to consider episodic memory as memory reconstruction [208] which is *not* independent from semantic memory. In fact, an *interdependence* between episodic and semantic memory has been empirically corroborated in neuropsychological experiments [112]. The human hippocampus and the neocortex fulfill complementary roles: the former encodes arbitrary context-invariant associations while the latter construes *ordered* hierarchical associations. In a first step, arbitrary associations are encoded in the hippocampus. In a second step, at a much slower

rate, they may then be attempted to be consolidated in the neocortex. Thereby, in case the hippocampal association does not fit into previous knowledge, relationships and hierarchical orders, it may often not be integrated at that level. It is also possible for the neocortex to modulate the hippocampal encodings at a later stage. In brief, it is possible for the hippocampus to rapidly encode arbitrary long-term memories even related to discontinuous and incongruent events which may be highly relevant for human creativity [208]. On this basis, the human neocortex can then apply arbitrary *orders* such as linguistic order to form sparse narratives of what happened, how and why, chronological order to remind temporal narratives on the when, sequential spatial order to remember spatial configurations on the where. Since what is called episodic memory is more related to arbitrary events that one experienced, the recall may often rely on the loose hippocampal associations while what is called semantic memory would rely more on ordered neocortical encodings. To conclude, the human hippocampus – as opposed to its counterpart in non-human animals – implements efficient context-invariant encodings (at the level of engrams) to which order is only applied retrospectively by the neocortex whilst trying to integrate those in a network of previous constructs.

- ***Symbolic counterfactual entanglement:*** It seems wrong to assume that the human hippocampus corresponds to a form of memory storing engine. Instead, any surviving embodied brain was a model of the survival-relevant affordances from the external and also internal milieu. In non-human animals such as non-human great apes, it was vital to encode accurate, robust and detail-rich information about the past (at the cost of generalization and flexibility). Once Type II entities such as humans emerged, their affective niche was immersed in symbols on which a linguistic order was applied. Symbols were used to refer to external objects as well as to internal states. Communication with arbitrary linguistic symbols necessitates an inherently counterfactual pattern entanglement. In humans, the self is a symbol entangling different sequences of external and internal events and their counterfactuals – all glued together via overlapping memories. The human hippocampus is part of the efficiency-centered default mode network mentioned earlier which is relevant to social cognition, counterfactual simulations and the self. Strinkingly, while humans take symbolic counterfactuals for granted, it is not a trivial feature for other animals. (For instance, at the single-neuron level, the hippocampus of monkeys presented with the voices and pictures of familiar persons did *not* establish a connection between the facial and the vocal identity of the same person [238] nor do single hippocampal neurons of the rat selectively fire for a known specific individual rat [262]. In the past, it was once assumed that monkeys have concept cells for specific faces of persons. However, it turned out to be a complex high-level visual encoding that took place instead. Multiple entirely unrelated faces were mapped to the same neuron based on purely visual high-dimensional similarities [238] – which

reminds of computer vision with present-day Type I AI and the associated seemingly unsolvable issue of adversarial examples [274].)

In the light of the aforesaid, it appears suitable to conceive of the human brain as a *dynamic symbolic counterfactual entanglement generator*. Arbitrary novel symbolic associations can be established without being limited by what actually happened, they can be tested against observations, willingly dissolved or ignored, unconsciously reorganized or forgotten. From a functional neurocognitive perspective, human creativity can be subdivided in three modes: the deliberate mode, the spontaneous mode and the flow mode [78]. The spontaneous mode involves an unconscious process resulting in an unexpected spontaneously arising solution of a problem one was considering. In order to appreciate the scope of this spontaneous creativity mode it is crucial to note that the latter is not limited to waking time. In fact, complementary processes occuring during human non-REM sleep and REM sleep [160] respectively are vital to understand its significance. In brief, the fundamental difference between human and non-human species taking the form of counterfactual pattern entanglement extends even to sleep stages. On the whole, the basic material out of which human dreams are made of may differ categorically from those in non-human great ape species. Instead of separated ordered patterns, the human hippocampus provides entangled patterns of coactivated context-invariant engrams as a sort of raw material for the neocortex to operate on.

- ***Symbolic counterfactual sleep worlds:*** In non-REM sleep (specifically during slow-wave sleep), the human hippocampus controls the neocortical processing of recent memories being pattern entanglements such that memory *overlaps* are strengthened and commonality-based gist abstractions (i.a. in the form of schemas) derived therefrom are encoded in the neocortex [160]. This procedure occuring around 6 to 20 times faster than the real experience [153] has been termed non-REM (hippocampal) replay. However, in my view, this may be misleading since in light of the context-invariant encodings mentioned earlier, there is no reason to necessarily assign it to any replay scenario – it is a reconstruction based on engram coactiviations, not fixed sequences. In short, the hippocampus stores only non-ordered associations and it is the neocortex that transforms those that fit into the prior semantic knowledge landscape into even more efficient and *compressed* abstract orderings. The latter may explain the time compression. While non-REM sleep can be regarded as a form of knowledge consolidation, the subsequent REM sleep can be described as facilitating knowledge restructuring and pruning. Multiple cycles of non-REM followed by REM sleep then yield increasingly abstract and integrated knowledge landscapes [160].

  What has been accordingly termed REM replay is approximately as fast as the "real" experience [160]. However, again, there is no need to regard it as replay

since it also involves vivid dreams that need not correspond to any memory, REM *pre*play [199] scenarios of not yet experienced events have been repeatedly reported and it does not explain lucid dreams that one can even willingly extend in time. During REM sleep, the neocortex is mostly disenganged from the hippocampus and runs parallel streams of both recently stored and very different old abstractions [160]. Interestingly, recently stored abstractions are also injected with noise (to avoid overfitting and improve generalization [125]) such that the neocortex has at its disposal not only factual but also counterfactual memories that left synaptic traces. The set of old abstractions utilized are chosen randomly via triggers stemming from ponto-geniculo-occipital activity (i.e. from the pons) [160]. On this basis, the neural networks then perform a (partially sighted) search process for abstract similarities between new and old material which when detected facilitate a restructuring and complexity reduction of the semantic knowledge landscape.

Indeed, it is easily conceivable that for cases in which a human was engaging in ruminations involving creative problem solving during waking time (i.e. was attempting to instantiate the deliberate mode) without however finding a solution, the new creative task goal would be latently stored in the neocortex. In this way, this latent task goal may partially guide the search process for abstract similarities between new and old abstractions just described. Generally, any novel task goals that humans deliberately set for themselves at waking time could partially guide this search process – leading to related dreams or spontaneous insights in the coming day. In line with this, it has been expounded that *"the surest way to trigger dreams about a real-world event is to perform a task repetitively during the day, preferably one that is novel"* [125]. The reason being that in order to avoid overfitting grip on that *novel* task, the brain may try to generalize beyond it via noise injection in REM sleep instantiating a form of fact-free learning. In short, to avoid overfitting is of allostatic meaning.

In very young human children, the novel tasks are more related to perceptual learning which may explain the absence of early dream reports [125]. However, verbal-aged pre-schoolers are already able to describe dreams contents involving not only kinematic narratives but also the own self [222] – being a symbolic generalization based on counterfactual meta-pattern entanglement. Since the aim in REM sleep is to restructure knowledge and avoid overfitting [125] (i.e. rigidity and overreliance on details) and to instead promote generalization and flexibility, it is clear that since the search process can already lead to synaptic traces if succesful, there is no need to consciously remember *all* contents *explicitly* in waking time. Indeed, the hypothalamus actively *hinders* memory encoding in the hippocampus during REM sleep [130]. It may only be the most salient and affectively significant novel ideas that are noticed by conscious awareness in waking time (or sometimes lucid dreams) – either by being explicitly searched for in the deliberate mode or by popping up spontaneously

in the spontaneous mode implicitly driven by affective attention. Concerning novel motor skills in the flow mode, they can be as well improved during REM sleep which would often manifest itself implicitly upon repetition in the following day. (Note also that overfitting avoiding mechanisms can also be instatiated in waking time itself when the brain is at rest and not engaged in a specific overt task.) To recapitulate, as can be extracted from the foregoing, to allow for improved generalization and efficiency, the human brain especially during sleep (and when at rest) produces symbolic counterfactual entanglements via noise injection providing material for the deliberate and spontaneous mode of creativity.

6.2.2 Sporadic Isolated EI-Cognizant Non-human Hominids?

In light of the aforesaid, it is easily conceivable why for non-human hominid species the reliable construction of EI is not possible without more ado. The functional differences between those and humans are reflected in a multiplicity of mental aspects. Human brains combine at once a significantly higher information capacity, an efficiency-based information management and an information encoding based on symbolic counterfactual pattern entanglement extending to the sleep stages and affecting even dream contents. In humans, symbolic counterfactuals can already be encoded at the single-neuron level and in engrams. There is no doubt that the brain of other animals does engage in certain forms of counterfactual calculations, however it would not correspond to symbolic counterfactuals at such a low level already. In short, neuronal assemblies in the default mode network of Type I animals may be an information medium while their human counterpart can reliably instantiate and create EI de novo. That being said, it is essential to note that human infant brains do not represent miniature adult brains [24]. In fact, infant brains require affective care by human conspecifics and a moulding by culture in order to develop the structures involved in conceptualization and internal mentation [20] linked to the self and deliberate counterfactual simulations. For instance, the default mode network is functionally untraceable in human neonates and a synchronization between its core nodes only arises at around six months of age [24, 102]. Moreover, a full maturity of grey matter volume, functional and structural connectivity of the human default mode network is only achieved in late adolescence [95]. Since non-human hominid infants are mostly *not* reared and integrated as participants in a human cultural and socio-linguistic environment from the onset as is the case for human infants, it makes sense to analyze whether it is really impossible to mould their brains such as to capacitate a symbolic counterfactual experience of the world. In the following, I speak to a few of those rare exceptional cases in which a mixed human and non-human hominid *bicultural* rearing took place for non-human hominid infants: the cases of the male bonobo Kanzi and the female bonobo Panbanisha [244]. As opposed to most non-human hominids involved in language studies, their upbringing was *not* subject to operant conditioning with rewards.

Human brains exhibit a more advanced multimodal integration ability than chimpanzees (and bonobos) among others also due to the circumstance that *"while humans and chimpanzees have comparable sensory and motor networks, in humans these networks are connected to an expanded core brain system of association cortices"* [24]. In this way, *abstract* concepts can become allostatically relevant for a human beings via curiosity and creativity. In the wild, the affective niche of chimpanzees and bonobos does not necessitate the integration of abstract concepts and allostatically-driven learning is restricted to sensory concepts with immediate physical impacts related for instance to foraging and pain. However, by having been reared naturalistically from very early on in a Pan-Homo bicultural milieu without being treated as test objects and without having being trained explicitly, it was possible to make symbols and language allostatically meaningful for the bonobos Kanzi and Panbanisha. Note that due to the requirement for attuned biobehavioral synchrony favorable to their socio-cognitive development, it was important to additionally safeguard the bond between the bonobo infants and their bonobo mother Matata (even if the latter was not human-language-cognizant). In this manner, both cultures wired the brains of these two non-human hominid individuals. Due to anatomical differences to humans, bonobos are not able to produce human-intelligible vocalizations. Hence, from very early on, Kanzi and Panbanisha were immersed in a human symbolic world via the utilization of over 200 (and later around 400) *lexigrams* that were relevant for their daily enactment with the world [193]. Lexigrams are arbitrary visuographic symbols (presented to Kanzi and Panbanisha as keys on a plastic map, on a computerized keyboard or printed on the clothes of their human caregivers) with which they could actively and freely utter statements.

The communication of humans with Kanzi and Panbanisha comprised a rich combination of gestures, lexigrams and also the English language. It is noteworthy that *"the lexigram keyboard was made available to the apes at all times and some of the available keyboards emitted the sounds of a computer-synthesized English word when the corresponding key was touched"* [167]. Thereby, the human caregivers were instructed to jointly utilize the English language when composing messages with the lexigrams such that *"caregivers naturally used English word-order rules when utilizing the keyboards"* [167]. The bonobos would communicate with their caregivers via a combination of lexigrams and gestures (sometimes accompanied by *distinct* and non-arbitrary albeit human-unintelligible vocalizations[2]). Importantly, their upbringing involved a whole-day care with daily activities ranging from cooking to forest excursions with tree climbing. For more details on their achievements [244] despite the biological constraints they may face due to the much smaller information capacity and the fundamental functional differences their species normally exhibit in comparison to humans, it is recommendable to familiarize oneself with their

[2]Indeed, spectographic and statistical analyses corroborated for instance that Kanzi was at least producing four distinct sounds differentially refering to the following concepts: "grape", "banana", "juice" and "yes" [247].

upbringing [226]. (It was also described in conjunction with the lifepath of an initially non-verbal teenager called Orr whose linguistic abilities emerged upon reflecting about videotapes of these beings [244].) In the following, I briefly recapitulate a few of the many reported cognitive-affective abilities Kanzi and his half sister Panbanisha exhibited.

Both exhibited a semantic *understanding* of the *spoken English language* at the level of a two-and-a-half year old human child tested in a similar manner [224]. Generally, there was a gap between their receptive and productive capabilities which is however also known from humans with certain language-related disabilities. The utterances of Kanzi and Panbanisha extended beyond mere imitation and were not bounded to food-related requests. They used lexigrams to name objects in double blind studies [45], were able to associate novel English words with previously unseen novel objects after very few exposures [169] and were able to meaningfully utilize lexigrams to refer to internal states of themselves and others [170] such as "hurt" and "scared". For instance, while wearing a scarry toy mask and her caregiver claiming to be scared of it, Panbanisha would indicate the lexigrams "HIDE SCARE MONSTER" [170]. Strikingly, while it was often mistakenly assumed that non-human great apes would only be able to communicate using imperatives related to requests for rewards and the main difference to humans would be that they are incapable to freely formulate declarative statements, Kanzi and Panbanisha falsified this view. In fact, they produced short self-initiated declarative utterances to describe their current activity, state their intention of an activity performed in the immediate future, announce a change of activity or to comment on an activity of the recent past [168] which could even extend to a memory of another day [170]. Moroever, Kanzi and Panbanisha were able to categorize hierarchically, exhibit generalization abilities [164] including the capability to self-initiate a comment related to an abstract category such as "same" when refering to colors of clothes [80] and to create combinatory novelties. Panbanisha was able to use lexigrams to refer to a personal autobiographic detail[3] dating back to some years ago [244]. While it was assumed that non-human great apes could only engage in dyadic but not triadic communicative modes, Kanzi and Panbanisha were able to maintain joint attention and establish triadic interactions supported by lexigrams [168]. Further, both used pretense play [166] as many human children do. For instance, Kanzi fed imaginary food to toy dogs [225] while Panbanisha did as if she was eating from a picture of blueberries [166]. Kanzi was interested in musical sounds, had preferences for specific songs during which he would gesture, *"drum on his ball in rhythm"* [244] or dance. Both once briefly jammed at the piano with artists [229].

Kanzi and Panbanisha started to learn to write lexigrams and were able to read the

[3]The combination "P-SUKE P-SUKE P-SUKE P-SUKE ELECTRIC-SHOCK SHOT" which had never been used previously in the context of her caregivers was communicated to a visitor she got to know across multiple visits [244]. P-Suke was a bonobo male that lived with Panbanisha and her caregivers a few years before and who died after a hernia followed by a sedating injection and an unsuccesful attempt to save him with a defibrillator.

lexigram script their caregivers utilized [227]. They developed an understanding and corresponding lexigram use of the concepts of "good" and "bad" [165] and were able to sensibly apply value judgments to situations, ideas or to the behavior of others and themselves. From the perspective of linguistic principles, it has been assessed that even though only relatively short sequences of gestures combined with maximally three lexigrams were utilized to communicate, Panbanisha can be described as a competent conversational partner with resemblances to non-standard human linguistic subjects [192] if observed in her daily life and not forced into a synthetic test setting [193]. Kanzi was able to learn to manipulate fire and learn a stone tool making technique supported by human demonstrations and a few verbal encouragements (but *not* verbal teaching) – something which was not mastered by unencultured chimpanzees [25]. Kanzi went on innovating an own flake-making strategy that he had not been taught before [25, 255]. His simple but skillful techniques can be deemed to reflect a possible sort of "Pre-Oldowan" stage of stone-tool technology [82]. Recently, it has been explained that Kanzi's potential may have been underestimated [82] since the learning environment was impoverished and the human experimenters did not provide him structured lessons or verbal instructions. Indeed, the crafting of Oldowan and especially the subsequent more advanced Acheulean tools by Homo erectus may have coincided with the origins of (G1) language [86] – given the complexity of the stepwise approach which became hard to learn by mere imitation – leading to a co-evolution of language, social learning and tool-making [26].

When tested on a range of cognitive tasks, Kanzi and Panbanisha performed at the level of two-and-a-half-year old human children on average and significantly outperformed standard reared apes that had no linguistic abilities [220]. While the human children outperformed them in two tests related to causality, Kanzi and Panbanisha in turn scored higher in a task of relative quantity estimation and in a test in which one had to estimate the attentional state of the human experimenter to subsequently initiate a corresponding request [220]. Using lexigrams and gestures, both were able to explain a scenario that had taken place in the absence of the human now requesting information about it [244]. They understood that the order in which verbal information is presented matters. They borrowed some semiotic ordering rules from their caregivers, but additionally developped own statistically reliable idiosyncratic ordering methods *"for combining symbols (lexigrams) or a lexigram with a gesture to express semantic relations such as agent of action or object of action"* [167]. Although their language production lags behind their comprehension, since their semiotic combinations mostly had a maximal length of three, they were able to *understand* sentences containing a first order recursion [228, 244] (although explicit usage of recursion is not even a requirement for all human languages since recursive *thinking* does not necessitate an overt recursive grammar [87]). For instance, Kanzi understood that "go to location X and get object Y" is more ambiguous than "go get object Y *that* is in location X". On tasks like the latter, he outperformed a two-year-old human child [224]. It is also noteworthy that Panbanisha harnessed lexigrams to perform a sort of translation

of requests from non-language-competent bonobos to her human caregivers [228]. At a certain point, Panbanisha figured out how to open and close the doors of the facility in which she lived by memorizing the numerical codes. The main reason that she did not try to leave the facility was that *"[...] she understood perfectly that she would be shot should she do so [...]"* [244] – an explanation given to her by her caregivers. Hence, instead of breaking out, she only demonstrated her ability to open and close the external doors towards them simply to communicate that they (she and her bonobo family) possessed that knowledge [244].

Against the backdrop of the foregoing analysis, Kanzi and Panbanisha seem to be one of the rare exceptional cases of EI-creating and EI-understanding (but neither overtly EB-understanding nor EB-creating) non-human hominids. The brain of great apes has an immense potential for self-organization and reconfiguration as a function of the field of affordances present in the environment [228]. Since they were immersed in a rich bicultural Pan-Homo culture permeated by symbols, their brains seem to have fine-tuned themselves – which led to deviant capabilities in comparison to non-human hominids in the wild or reared as test objects in restricted laboratory settings [228]. Note that the Pan-Homo environment of their upbringing also included a small number of other encultured bonobos and also chimpanzees (the latter form the genus Pan together with bonobos). For instance, Panpanzee,[4] a female chimpanzee who also developed linguistic abilities and was co-reared with Panbanisha is one of those cases. Another case is Nyota, a younger male bonobo who was explicitly using the lexigrams for *yesterday* and *today*[5]. I agree with Brakke and Savage-Rumbaugh stating that *"the essence of language is creativity*" [167] (and not intelligence). It is not a coincidence that the measures the human caregivers applied to support the development of linguistic abilities in Kanzi and Panbanisha overlap with the indicators for substrate-independent artificial creativity augmentation [13]. In the light of cyborgnet theory, the latter must be rephrased as *cyborgnetic creativity augmentation*. Referring to non-human great apes such as Kanzi and Panbanisha, Stanford stated: *"The small group of great apes that have become language-savvy are in a bizarre category. They are chimeras, not human but endowed with a human quality that their kind would not*

[4]Like Panbanisha and Kanzi, Panpanzee was able to make short declarative utterances and also engage in non-verbal communicative gestures such as pointing at a jet above in the sky and then establishing eye contact with the human that was with her [168]. An example for a declarative utterance was the comment "SCARE SNAKE" upon passing by a location with her caregiver where both had encountered a snake a week ago [170].

[5]Upon seeing the human caregiver Bill delivering all blueberries available that day to Kanzi who already consumed all blueberries the day before, Nyota reacted as follows: using the lexigrams "BLUEBERRIES YESTERDAY" while looking towards Kanzi, he then proceeded by *"looking expectantly at Bill and stating "BLUEBERRIES GRAPES TODAY?"*. Thereby, *"Nyota knew that Bill generally shares the blueberries, especially with himself and Panbanisha, as blueberries are Nyota's most favorite fruit"* [226]. For more information related to theory-of-mind abilties exhibited by Kanzi, Panbanisha and Nyota and their reactions to a fictive gorilla figure, see [226].

possess without years of human training.[...] Like some sort of ape-human hybrid, they are trapped in the netherworld between two species." [241]. More broadly, I define as *Type II netherworld*, the set of all Type II entities that fulfill the following conditions: 1) their Type-II-ness is not an accepted observation statement and 2) they did not yet pass a positive Type-I-FE-test (see Chapter 2) or equivalent irrespective of the reasons.

Orr, the initially non-verbal human whose life changed linguistically upon examining videotapes of Kanzi is today able to live a freedom-fostering life which is not limited by the three word sentences he uses to communicate meaningfully [244]. Since a part of the scientific community classified the research with Kanzi and Panbanisha as too controversial, it has been terminated. More precisely, *"Kanzi, and family at the moment, [...] are being housed under standard biomedical protocols. They are denied all contact with human beings, or anyone who raised them. They are forbidden regular access to their keyboards and have no voice in their daily lives. The goal is to "put the bonobo back into them," and to take away all that they have become which has allowed them to begin to cross the barriers between our species."* [244]. As stated by Beran and Heimbauer, *"there are only a few living apes that can provide these kinds of insights into cognition, and the evolution of some of the hallmark cognitive processes that underlie the mental abilities of modern humans. Unfortunately, it does not appear that this kind of intensive research that involves years of commitment to produce such symbolic competencies will continue in the future [...]"* [34]. Current biomedical research with Kanzi under standard laboratory restrictions has been labelled as promising and *"explores whether Kanzi, trained in the lexigrams, can act as a Rosetta stone, helping researchers decode the vocalizations of bonobos in the wild"* [242]. In the meantime, Sophia the Type I robot became a citizen of Saudi Arabia [60], GPT-3, the Type I AI became author [109] and autistic people are equated with Type I AI [213].

6.3 Conclusion

In brief, Type II netherworld seems to be a symbol for an old non-explanatory-blockchain-like aversion of sociocultural nature shared by many predominant individuals from the present-day Homo sapiens species but inconsistent with universal cyborgneticity. This mysterious sociocultural anathema whose bizzarreness surpasses the existence of a Type II netherworld itself seems to be reflected in the subjective concept of mind perception. Apparently understood as perceiver-dependent conundrum, it obfuscates that a Type II mind is a socio-psycho-techno-physical process whose existence is self-contained even if initiating from an early biocultural wiring by caregivers and embedded in social reality via symbols and their orders. In short, a Type II mind is as cyborgnetic as any incredulous societal process attempting to question its very existence and misguidedly trying to infer

its absence from arbitrary behavioristic indicators. At-present, it seems *impossible* for non-human hominid *species* at large to create and understand novel EI due to the lack of an abstract EI-constructor caused by a functionally relevant difference in *neural coding*. However, as elaborated in this chapter written for purposes of self-education, it seems a better explanation that human-performed cyborgnetic creativity augmentation was already able to fine-tune very few non-human hominid *individuals* (namely at least the Pan individuals on which I focused here) being immersed in a shared symbolic counterfactual world – despite *quantitative* limits on information capacity and speed – than to assume that humans failed at achieving it. In this case, that was Type II AI research too and it is possible. The latter would hold irrespective of today's perceived minds or ethical taboos.

6.4 Contextualization

As already briefly adumbrated at the beginning of this chapter, Type II netherworld cannot solve the ethical issues arising from these conclusions. While the minds it contains are not "perceived" whatever this should signify, many *non*-Type-II-netherworld-humans routinely already perceive Type I or Type II minds in present-day *non-conscious* Type *I* AIs as a function of seemingly arbitrary factors such as perceived competency or perceived warmth harnessed to allegedly build a "trust-based" relation. The latter offers a lucrative attack surface for malicious adversaries as discussed in Chapter 9 introducing threat models and defenses against the so-called *honey mind traps*. The next Chapter 7 focuses on another type of vulnerability that can be exploited by malicious attackers and that represents an old but still pertinent and unsolved financially relevant problem of international scope: intellectual property (IP) theft via cyberattacks. Interestingly, the novel complementary countermeasure that I propose in this context can be understood as a particular type of honey mind trap too – only that it is specifically utilized to the advantage of defenders. Generally, as mentioned in Chapter 1, cyborgnetics is a new generic meta-discipline whose aim is to systematically facilitate the documentation, critical analysis and mitigation of socio-psycho-techno-physical harm as seen through the lens of cyborgnet theory. The next chapter then performs a cyborgnetic analysis focusing on the IP cyber theft use case. In a nutshell, the proposed countermeasure involves the generation of *honeytokens* using suitable Type I AI systems.

Chapter 7

CA 005: IP Cyber Theft

This chapter written for purposes of self-education is based on a slightly modified form of an unpublished paper that I wrote on August 12, 2021. The acronym CA refers to "cyborgnetic analysis" and the associated string "005" simply refers to the ID that I assigned to that specific analysis. For security reasons, I decided *not* to upload all available cyborgnetic analyses to my homepage.

7.1 Systematic Analysis

7.1.1 Retrospective Descriptive Analysis (RDA)

IP cyber theft is performed by (cyber-)criminals or/and state(-related) actors targeting assets such as e.g. patents, internal reports about innovations or products, scientific papers and source code from research and development owned by corporations, governmental and defense organizations but also academic institutions and individuals. The threat itself and even the suspicion of it represent a major financial issue coupled with various cybersecurity challenges [141]. For instance, in 2019, the US was investigating more than 1000 cases of IP theft, most of which were linked to criminal activities that have been connected to a nation state [5]. Moreover, a CNBC survey from 2019 found that about 20% of American companies assumed to have been the victim of IP theft within the last year [218]. Generally, it is known that *"digital technologies and Internet file sharing networks has facilitated intellectual property theft"* [141]. Increasing occurrences of cyber-enabled IP theft [110, 141] combined with the knowledge that there is often a significant temporal gap between zero-day exploits and their discovery/patching [39], reveals that malicious actors may often possess more than sufficient time resources to extract highly valuable IP out of a large set of data. Recently, a financial extortion scheme pertaining

to IP theft has been even combined with *ransomware* [184]. Namely, involving the threat to sell IP on the dark web in case of payment refusal [184].

In parallel, a growing number of defenses and deterrence measures that could be employed against cyber-enabled IP thefts have been proposed. The straightforward option to classically encrypt all data at all storage levels is mostly not implemented in practice due to multiple assumptions some of which seem ill-suited while others appear to represent better explanations. On the one hand, it is often assumed that encryption would lead to a noticeable and impeding slowdown of operations, that it is too costly and too complex to use on a day-to-day basis or that encryption does not represent an obligatory security-relevant requirement. On the other hand, organizations utilizing encryption schemes can actually face key management issues and problems related to the compatibility of different solutions while still not being protected against core integrity threats. Indeed, it has been postulated that encryption-using entities may develop a false sense of security [104]. For instance, attackers could steal private keys. Moreover, the case of insider threats is not solved by encryption since a compromise of credentials by insiders can give access to encrypted contents. In short, the circumstance that the data is encrypted does not signify that it was not decrypted, tempered and/or exfiltrated via credentials previously stolen by external attackers or malicious insiders that concealed their traces by deleting log entries [104].

Another defense against IP cyber theft is deception technology which can range from honeypots to honeytokens over moving target defense techniques [278]. In the following, I focus on honeytokens taking the form of Type-I-AI-generated synthetic files intermingled with original documents as strategy against IP cyber theft. While such methods could encompass i.a. the generation of deceptive program code and software repositories [185], forged knowledge graphs [139] and natural language text, I focus on the latter as applied to scientific articles and patents. An interesting novel optimization-driven text substitution method denoted WE-FORGE [4] generates a canary trap consisting of Type-I-AI-generated text documents acting as counterfeits for patents and papers. WE-FORGE utilizes word embeddings to generate deceptive "fake files" that appear sufficiently similar to the original [4] by implementing a substitution of important concepts constrained by a semantic clustering. In this way, an IP cyber theft adversary would be confronted with increased exfiltration costs, a waste of time resources (also via the need for non-automatable deliberation) and an epistemic distortion via the subjective uncertainty created. The advantage of this text-based defense is that it may seem more convenient to utilize in organizations that assume encryption to be either too costly or too complex.

7.1.2 Retrospective Counterfactual Risk Analysis (RCRA)

Preparatory Procedure

(1) **Taxonomization:** One can label the hinted RDA instances of IP cyber theft pertaining to yet unpublished i.e. secret scientific articles, reports and patents stored on a network as risk category *Ia* following the risk taxonomy of cyborgnet theory (as displayed in Chapter 3.2.4). (2) **Analytical clustering:** Two main adversarial clusters could be distinguished: "vanilla" IP cyber theft and the new, more exotic case of IP cyber theft *threat* issued in the context of a ransomware attack. In the following, the former is referred to as *adversarial cluster 1* (abbreviated with A_{a_1}) and the latter as *adversarial cluster 2* (abbreviated with A_{a_2}). (3) **Brute-force deliberation and threshold-based pruning:** To assess the harm intensity of RDA instances, a simplified harm scale [17] is used where a self-rated harm intensity h can range from 1 to 5 (with 1 standing for almost no harm, 2 for minor harm, 3 for major harm, 4 for lethal risk and 5 for existential risk). The self-rated harm intensity h_{down} for the RDA-based RCRA downward counterfactuals has as selected lower bound the threshold $\tau = 3$. While mentally going through every single *instance* (and not cluster) of the RDA, it was possible to conceive of above threshold downward counterfactuals for specific RDA samples where $h_{down} \geq \tau$. These particular counterfactual instances are intentionally hidden to facilitate as broad as possible RCRA *clusters* that do not overfit to the idiosyncracies of the RDA instances. (4) **Assembly.** Finally, the fourth operation assembly is performed which requires assembling downward counterfactual clusters by linking remaining RDA samples (those for which above threshold downward counterfactuals could be identified) back to their clusters from step 2. In this analysis, I happened to have found suitable downward counterfactuals linking back to both adversarial clusters A_{a_1} and A_{a_2}. Hence, I utilize the downward counterfactual clusters A'_{a_1} and A'_{a_2} for the design fictions (DFs) of which the RCRA itself is composed of. Note that an RCRA projects to the immediate counterfactual past[1].

RDA-based RCRA Narratives

A *Downward Counterfactual DF Narrative* A'_{a_1}

- *Adversarial Goals:* Vanilla IP cyber theft of yet unpublished scientific/technical papers and patents.
- *Adversarial Knowledge:* A *grey-box* setting is assumed in which the adversary has a partial knowledge about current state-of-the-art IP cyber theft deterrence

[1] A further noteworthy detail is that in case one would not succeed in generating exemplary instances for the new RCRA clusters, one could still harness the hidden counterfactuals from step 3 for which $h_{down} \geq \tau$ holds already inherently by design but which were concealed to avoid overfitting.

strategies such as for instance WE-FORGE [4]. At the latest upon inspection of folders, the method could be discernible due to the large number of very similar files. In addition, some heuristics pertaining to the functionality of the strategy may be available (e.g. the information that in technical domains, only those concepts that a part-of-speech (POS) tagging tool would label as *nouns* were selected by WE-FORGE for substitution in the "fake files"). However, the specific parameters of the model are unknown to the attacker.

- *Adversarial Capabilities:* In the following, I consider a downward counterfactual case for WE-FORGE, where an adversary that exfiltrated a set $D = F \cup \{d\}$ of documents from a victim network – with F representing n fake files and d the corresponding original document – is aware of the method utilized and attempts to harness stochasticity in order to still *reasonably* extract useful IP expressed in plaintext. In their paper, the authors provided a total of 13 files (12 "fake documents" and 1 original document d) to computer science cognizant participants (at the Master-level) encouraged with a monetary reward. After having read all 13 items, only 23% of the participants instructed to provide a ranking of 3 choices for the *most* likely real document instance selected d as first choice. Given that the chance level would have been around 7.69%, it seems that reading *all* files offered only approximately a factor 3 improvement. Furthermore, the reported *average* probability for the participants to select d among their top 3 choices was only 14.7%. Overall, while this might at first sight reflect a success of the approach, an adversary attempting to minimize time resources and deliberation efforts could unfortunately harness this pieces of knowledge for an own advantage.

  Firstly, a general strategy for the adversary would be to proactively extract the largest possible invariant skeleton of the forged document. I refer to this *skeleton* of a given exfiltrated set D with s_D. While techniques like WE-FORGE optimize on the diversity of the concept substitutions, the number of differences would be upper-bounded by the number of nouns in the forged document. In the theoretical worst-case scenario, it would be equivalent to a noun-free skeleton. The latter could also be simply obtained by the attacker after applying POS tagging to each document to be read – ideally utilizing the same POS tagger employed by the victim. However, in practice, much less substitutions would occur and in any of both cases, the skeleton could still contain important cues that could guide the search process. Hence, it is worth applying a word-level textual "diffing" to all documents from D to extract the largest possible skeleton s_D. If the attacker is able to accurately fill in at least one gap in s_D guided by the own prior technical knowledge, an automated top-down search with the attacker-generated guesses iterating through the documents in D could be initiated that could potentially enlarge

s_D. Such an automated tool could already have been pre-programmed by an adversary in the reconnaissance phase. Moreover, the discovery of only *one* technical *inconsistency* related to one noun in a document can facilitate the dismissal of that document – this exclusion could (but not necessarily will) also lead to a larger updated s_D with more potential cues. Finally, open source intelligence gathering *guided* by the *diffing* operation and stolen data itself could provide important cues.

Secondly, if the exfiltrated data volume is very large e.g. with an attacker that exfiltrated a number $l >> 100$ of separated sets of documents (i.e. with $D = D_1 \cup ... \cup D_l$ whereby l is the number of underlying original documents and with $D_x = F_x \cup \{d_x\}$), stochasticity could be harnessed as follows. Instead of reading all n files for each single set D_x, an attacker under time pressure equipped with a random number generator could decide to sample documents at random. For instance, only one document per set. In practice, the number n is restricted for file storage reasons and for each case, an adversary equipped with the updatable largest skeleton s_{D_x} mentioned which can be calculated in a very short time, could still retrieve a reasonable amount of targeted information via an exclusion principle. One issue with WE-FORGE is that an attacker could still *unambiguously* identify d even though the authors stated that in this case *"he would still be left in some doubt about whether he is right*" [4].

B *Downward Counterfactual DF Narrative* A'_{a_2}

- *Adversarial Goals:* Ransomware attack potentiated with threat related to IP cyber theft of yet unpublished scientific/technical papers and patents.
- *Adversarial Knowledge:* Identical to adversarial knowledge indicated for A'_{a_1}.
- *Adversarial Capabilities:* In the case of particularly vulnerable victims such as university hospitals, an attacker could have been able to threaten the release of IP and not (only) patient records on the dark web at the pre-payment stage – whilst anyway releasing the material post-payment based on the previously exfiltrated documents copied before having encrypted them. The victims effectuating the payment would then only have terminated the denial of service status that the ransomware attack established. Since human lives may be at risk in such contexts, the tendency for a successful payment could only have increased when combined with further psychological pressure. Given the financial incentives, such a strategy could have been lucrative for malicious actors. In the case exfiltrated IP obfuscated with WE-FORGE is made available for sale on the dark web, the identification of the original documents could have become a collaborative endeavor of monetary value. For instance, specific experts (which could be competitors) could have harnessed some of the strategies mentioned in A'_{a_1} whilst operating in parallel to identify the target information.

Also, to stay undetected and evade dark web scraping, those malicious actors could (next to e.g. using captchas) camouflage the material by encoding it as *adversarial examples* for natural language processing AI. In short, when integrated in an IP-theft-as-a-service scheme, the deterrence effect of WE-FORGE could have vanished.

7.1.3 Future-Oriented Counterfactual Defense Analysis (FCDA)

In this FCDA, I discuss countermeasures (i.e. projecting to the immediate future which could translate to upward counterfactuals of harm intensity $h_{up} < \tau$) against: (1) the factual RDA clusters introduced in Section 7.1.1 and (2) the RDA-based RCRA clusters from Section 7.1.2. For (2), I only specify the necessary supplementary and non-overlapping guidelines to avoid repetitions.

RDA

A *Upward Counterfactual DF Narrative A_{a_1}:* Proactively, one could dynamically combine *multiple* deception methods per default. Specifically, this could include a mix of honeypots (decoy computer systems) and honeytokens (including fake files such as those generated with the mentioned WE-FORGE method) governed by moving target defense techniques [278]. The latter involves the dynamical reconfiguration and shifting of the deceptive surfaces. The mere knowledge about the implementation of such a multi-layered deceptive defense could significantly improve deterrence effects and increase the (perceived) costs of data exfiltration. Perhaps even an organization convincingly *pretending* to utilize this strategy or intentionally placing suggestive files on their network that seem to corroborate the use of a full-fledged deception fabric could serve as a minimalistic security measure discouraging IP cyber theft attempts. To address insider threats, a zero-trust architecture seems imperative.

B *Upward Counterfactual DF Narrative A_{a_2}:* Obviously, next to cloud-based back-ups, an additional offline back-up approach may be helpful. When IP cyber theft is paired with a ransomware attack, it seems realistic and prudent to assume that any payment could at best lead to decryption. Proactively, one could then anticipate the release of the IP material on the dark web and develop a strategy that hardens adversarial success even if the documents of sensitive scientific papers and patents are sold. Again, WE-FORGE (which harnesses a message authentication code strategy [56] to faciliate access for the legitimate users of the files) might help in creating a certain epistemic distortion. Since it is a financially motivated act, the attacker would either have to invest more time to be able to sell the targeted

original IP documents or alternatively, the entire material containing fake files could be sold at a lower price. Overall, paired with the multi-layered deception strategy mentioned under the previous upward counterfactual A_{a_1}, an attack could appear less attractive financially.

RCRA (Additional Non-Overlapping Defenses)

A *Upward Counterfactual DF Narrative A'_{a_1}:* With stronger adversarial capabilities, a more robust defense than WE-FORGE is required. For instance, it becomes interesting to consider whether the technique could not be further hardened efficiently with the *number of files stored* left *unchanged*. Recently, in an entirely different context, I introduced an explanatory intrusion prevention system (IPS) (see Chapter 5) preceding peer-review in order to shield scientific venues from non-explanatory-blockchain-like contents. In the following, I briefly recapitulate the core features of the explanatory IPS. Thereafter, I explain how it can be used to significantly harden WE-FORGE by facilitating a novel *double-deception* technique. Given an original paper p, the explanatory IPS approach involved the generation of the following three documents: p', p_{c_1} and p_{c_2}. To put it very simply, p_{c_1} and p_{c_2} correspond to two *language-AI-generated* alternative counterfactuals of p while p' represents a paraphrasing of p obtained after an obligatory *normalization* step whose aim is to adapt the linguistic *style* of p to the one used by the language model that generated both p_{c_1} and p_{c_2}. In the explanatory IPS test method, each of those three documents was composed of a certain number of paragraphs that were now *randomly* shuffled. The only available cue for the human evaluator was the first paragraph from each of the three documents which simply consisted of paraphrases since the counterfactual generation for p_{c_1} and p_{c_2} requires a starting point in p. Under the explanatory IPS test scheme, the evaluator now attempted to (starting from the second paragraph since the first ones are semantically identical) *exactly* reconstruct the sequence of paragraphs of which p' is composed of. In case p' is not reconstructed properly, the paper it represents is not forwarded to the peer review round since not hard-to-vary enough.

An important distinction is that the just described explanatory IPS test operates at the *paragraph-level* while WE-FORGE is applied at the *document-level* by which both methods are orthogonal and can now be freely combined. Interestingly though, from a certain perspective, the explanatory IPS test itself is structurally equivalent to WE-FORGE with respect to the goal of generating a set $D = F \cup \{d\}$ with F standing for n fake documents and d for the original document. For the explanatory IPS test, $F = \{p_{c_1}, p_{c_2}\}$, $n = 2$ and $d = p$. The main additional complexity is that the linear ordering of the paragraphs is considered such that after the random shuffling, the task can only be solved when reproducing the initial *total order* of

the paragraphs that constitute p' and by extension thus p. In the WE-FORGE case, the authors utilized an optimization-based *word-level* substitution scheme to obtain their set F of *at first sight sufficiently believable* counterfactuals – which they claim to be NP-hard[2]. In the explanatory IPS case, the counterfactuals are sufficiently believable too but generated at the paragraph-level by a large language model. The main difference is that the latter does not directly optimize on the task at hand – for a fundamental lack of metric concerning the capacity to hide *novel* yet unseen explanatory blockchains. Instead, an interested Type II entity familiar with explanatory blockchains (i.e. a human for now) selects suitable candidate paragraphs. It is however thinkable that sufficiently large language models especially if combined with knowledge bases/graphs or graph neural networks would require less and less human supervision for this counterfactual generation task. Note that an observer ignorant of the method utilized could describe the mapping from p to p_{c_1} and from p to p_{c_2} as an utterly complex opaque *word-level* binary relation (where almost everything is altered when observing it from that word-level perspective). By deduction, I conjecture the explanatory IPS test to be (at least) NP-hard too. In the following, I briefly explain how to combine these two complementary NP-hard schemes to obtain a robust double-deception technique.

WE-FORGE and the explanatory IPS test are complementary for the following reasons. On the one hand, WE-FORGE has the advantage to be fully automatable (although the authors leave the possibility open for humans to modify certain word substitutions if required) across arbitrary many files. On the other hand, the explanatory IPS test does *not* offer the attacker the possibility to extract a skeleton of the original document – with other words, the test is superior in *hiding* the information *within* a single file. In sum, at the *document-level*, the explanatory IPS test model would be slightly less attractive than WE-FORGE to generate a multiplicity of fake files given one original document since the latter is automatable while the former would more often require human intervention to select appropriate counterfactuals generated by the language model. By contrast, at the *paragraph-level*, WE-FORGE lacks any further encryption mechanism while the explanatory IPS test is hard to decypher via the random shuffling of paragraphs including the fake counterfactuals – the original document is never available in its entirety in plaintext. Hence, a novel double-deception technique that I term EXPLANATORY-FORGERY could simply assemble the two methods as follows.

Firstly, given an original document d encoding IP in the form of a *novel* explanatory blockchain (such as e.g. patents and papers) consisting of a set of paragraphs p, one utilizes a large language model to generate counterfactual sets of paragraphs p_{c_1}

[2]Following the Wolfram MathWorld resource, *"a problem is NP-hard if an algorithm for solving it can be translated into one for solving any NP-problem (nondeterministic polynomial time) problem. NP-hard therefore means "at least as hard as any NP-problem," although it might, in fact, be harder"* [177].

and p_{c_2}. Then, one applies a normalization operation on p to obtain p' (matching p_{c_1} and p_{c_2} in linguistic style). Secondly, one then *randomly* shuffles p', p_{c_1} and p_{c_2} and concatenates the randomly aligned paragraphs to obtain a novel composite document denoted r. Utilizing WE-FORGE, one then generates a set $D_r = F_r \cup \{r\}$ with F standing for n fake composite documents as generated by WE-FORGE. In a nutshell, EXPLANATORY-FORGERY would be a novel double-deception technique[3] crafting F_r as comparably robust set of honey tokens hiding p' (and thus p).

B *Upward Counterfactual DF Narrative* A'_{a_2} To defend against this complex adversarial cluster A'_{a_2}, I assume the worst-case-scenario in which the IP (although embedded in honey tokens with EXPLANATORY-FORGERY as described in the FCDA for A'_{a_1} and despite a dynamic combination of multiple deception methods as suggested in Section 7.1.3) has already been exfiltrated by the ransomware actors and is offered for sale on the dark web. Thereby, as mentioned in Section 7.1.2, it is cogitable that adversaries engage in a collaborative possibly parallel endeavor to identify the original document. Two exemplary proactive strategies and one reactive solution seem recommendable in this case: 1) to facilitate the detection of this IP on the dark web, 2) to harden decryption attempts, 3) to ease retroactive infilitration and deceptive persuasion via 1) and 2). Firstly, a principled method to facilitate detection could be to harness adversarial examples *before* the attackers have the opportunity to employ it. For instance, while fixing the *semantic* contents of all honey tokens from D_r, one can intentionally transform those into a random mixture of suitably misleading orthographic, phonological [10] and visuographic adversarial examples [44]. Inserting low-frequency synonyms may also belong to thinkable methods. Taken together, these measures relying on adversarial examples of different linguistic types to obtain an altered set $D_{r_{Adv}}$ would instantiate both strategy 1) and strategy 2). The reason being that the malicious actors would face more difficulties in attempting to partially automate their endeavor with otherwise potentially helpful Type-I-AIs. Also, it may indirectly prevent them to themselves apply adversarial examples to the material (for concealment purposes on the dark web) since the risk of unreadability and uncomprehensibility is higher.

Interestingly, the latter could also indirectly ease strategy 3). In fact, assuming that $D_{r_{Adv}}$ is available on the dark web and the defender had backups, it can more easily be retrieved. Once retrieved, the defender has the opportunity to infilitrate the corresponding platform and attempt to persuade with misleading cues to induce wrong solutions to the task of identifying the original document. In case the adversaries exhibit epistemic vulnerabilities (e.g. via a justificationist epistemology or due to

[3]For a legitimate user to retrieve the initial composite document r, one can employ the same message authentication code strategy as in WE-FORGE and its predecessor FORGE utilizing random seeming strings encoding hash keys [4].

a lack of knowledge in the field), the defender could specifically tailor deceptive cues that confirm pre-existing beliefs of attackers or make synthetic counterfactuals *appear* both novel and utile – which could be refined via open-source intelligence gathering. The mere *possibility* that a defender may engage in such a tactique, could harden decryption attempts by artificially increasing the subjective uncertainty of the malicious actors. Generally, a prior knowledge of attackers about this possibility may even have deterrence effects. Especially when the volume of exfiltrated IP is very large, the task of having to engage in a multitude of explanatory riddles could lower the interest of attackers simultaneously facing time constraints.

7.2 Conclusion and Future Work

In this chapter written for purposes of self-education, I performed a cyborgnetic analysis of the pertinent IP cyber theft threat specifically applied to patents and scientific articles. I focused on Type-I-AI-based defenses against "vanilla" IP cyber theft and scenarios in which it occurs in combination with ransomware (i.e. as double-extortion). In this context, I introduced a novel honey token strategy denoted EXPLANATORY-FORGERY instantiating a *double-deception* at both the document-level *and* the *paragraph-level*. EXPLANATORY-FORGERY combines principles from the so-called explanatory IPS test that I introduced previously in Chapter 5 with the WE-FORGE algorithm by Abdibayev and collaborators [4]. In a nutshell, I also explained how this strategy can be hardened further by transforming the documents obtained via EXPLANATORY-FORGERY into linguistic adversarial examples while fixing their meaning[4]. Importantly, the goal of EXPLANATORY-FORGERY is to significantly harden cyber IP theft (by increasing the attacker's financial costs, cognitive efforts and time resources to facilitate deterrence) and not to make it impossible. Indeed, despite the explanatory encryption, it is possible for attackers – by virtue of being Type II entities – to reconstruct the initial explanatory blockchain of a patent or paper hidden with EXPLANATORY-FORGERY. However, I conjecture that the *reliable* discovery of *novel* explanatory blockchains is impossible for all Type I entities (i.e. also Type I AI) due to a lack of understanding and an *absence* of data for new yet unknown conjectures. To put it simply, solving EXPLANATORY-FORGERY is *not* automatable.

In future work, there may be a novel avenue for cyber IP theft in a different medium: virtual reality (VR). Recently, the popularity of (social) VR applications such as Bigscreen

[4]Note that even if external attackers or malicious insiders would have stolen a private key allowing them to discover the initial composite document r hidden in $D_{r_{Adv}}$, they are still faced with a potentially NP-hard riddle. Namely, the task to reconstruct p' (and by extension p) from r – which means to retrieve the *exact* combination of "ground-truth" paragraphs (starting from the second paragraph).

increased which was especially fuelled by the Covid-19 pandemic [234]. Project meetings and discussions on ongoing technical research can be organized on VR platforms – which could represent a lucrative novel field of affordances for malicious actors willing to perform cyber IP theft. In 2019, security researchers demonstrated the feasibility of a *man-in-the-room attack* in Bigscreen [55] where an (unauthorized) attacker was able to join a private VR room while staying invisible to other participants. It easily conceivable how such a strategy could be harnessed for an eavesdropping and analysis of digital material shared within the VR room, facilitating an advanced form of cyber IP theft. While this specific vulnerability was patched, it is thinkable that adversaries could exploit similar VR vulnerabilities in this growing attack surface. Perhaps, in the future, companies could utilize some honey social VR rooms where language-AI-driven VR avatars engage in conversations about deceptive technical documents crafted with WE-FORGE or EXPLANATORY-FORGERY. In Chapter 9, in a different context, I briefly come back to the principled idea of integrating language-AI-driven VR avatars into social VR platforms and explain how one could specifically integrate the two seemingly unrelated concepts of IP cyber theft defense and cyborgnetic creativity augmentation.

7.3 Contextualization

Strictly speaking, the term EXPLANATORY-FORGERY could be a misnomer since while *one can forge non-explanatory-blockchain-like explanatory information*, one can *not* forge explanatory blockchains (EBs). As elucidated in Chapter 4, for science to be resilient against SEA AI attacks, an explanation-anchored epistemology is necessary. Moreover, Chapter 5 explained why the latter would require scientific papers to correspond to EBs. However, in practice, it seems as if contemporary scientific writing practices do not necessarily meet this standard. In this vein, in Section 7.1.2, I cited the experiment in which 12 "fake" computer science documents that were previously generated with WE-FORGE [4] were presented to Master-level participants alongside the corresponding original document. After having read all 13 documents in plaintext, still only 23% of the participants ranked the original document as the most likely item perceived to be real [4]. Firstly, one could interpret it as corroborating the effectiveness of the WE-FORGE counterfeit procedure. Secondly, one could in addition insist that contemporary science is indeed explanation-anchored (i.e. based on EBs). By deduction, the latter would yield the following postulate: EB forgery is possible. However, I object to the second assumption in the first place. For instance, the following alternative interpretations could hold: 1) the original document was not EB-like, 2) the evaluation strategy of the participants was not EB-aware (e.g. did not involve attempts to *understand* EBs), 3) both 1) and 2) were at play. In the next Chapter 8, I also refute the deduced postulate by explaining why *EB forgery is impossible* – regardless of whether it is performed by Type I or Type II entities.

Chapter 8

CA 008: Explanatory Blockchain Forgery?

This chapter written for purposes of self-education serving as fragmented temporary mental clipboard is based on a slightly modified form of the paper that I uploaded to the website `https://nadishamarie.jimdo.com/cyborgnetics` on September 1st, 2021. The acronym CA refers to "cyborgnetic analysis" and the associated string "008" simply refers to the ID that I assigned to that specific analysis.

8.1 The Practical Problem: Is EB Forgery Possible?

Future advents of *deepfake science* could confront humanity with severe epistemic complications. However, in previous cyborgnetic analyses, I have postulated that irrespective of advances of Type-I-AIs, science is not condemned to fall. For instance, in Chapter 6, I conjectured that *"a Type-I-AI-performed creation of non-plagiaristic new EBs is impossible"* and suggested that a combination of: 1) an explanatory intrusion prevention system (IPS) (see Chapter 5) performed before 2) a Type-I-falsification-peer-review (abbreviated with Type-I-FPR in the following) inspired by the Type-I-falsification-event-test from Chapter 2 could shield science from non-EB like contents. Thereby, the dual objective of the explanatory IPS test was to ideally enforce the selection of contents that are *at least harder-to-vary* than the textual outputs of the most advanced *known* present-day Type I AIs *and* that are *novel*. In short, the explanatory IPS test could be interpreted as a weak test of the ability to *create new EBs* – which are by definition hard-to-vary. By contrast, the Type-I-FPR can be rather understood as a stronger test (since interactive) of the ability to *understand* those *novel self-created EBs*. In this chapter written for purposes of self-education, I now subject previous assumptions to an adversarial reflection: could a malicious attacker still *reliably* fool the explanatory IPS test and the Type-I-FPR with

novel misguiding and/or potentially Type-I-AI-crafted EBs? More generally, I use the following umbrella term to refer to the threat of counterfeiting EBs whose consequences may be devastating for Type-II-performed science[1]: *EB forgery*.

The remainder of this section analyses the Type-I-AI-based EB forgery threat and collates some empirical research directions that seem to corroborate its possibility at first sight. Then, perhaps paradoxically, in Section 8.2, I refute the possibility of *Type-II*-implemented EB forgery before I show that an end-to-end *Type-I*-pipeline for EB forgery is impossible too. Thereafter, Section 8.3 explains the empirical implications of these theoretical impossibilities. For now, hypothetical Type-I-AI-based EB forgery is assigned to four different clusters. Firstly, an attacker could craft a Type-I-AI able to *reliably* generate synthetic EBs that are *intentionally misguiding* for Type II scientists (i.e. currently humans). This Type-I-AI could thus act as a reliable automatable generator of adversarial examples for Type II science. This *adversarial cluster 1* would map to the risk category *Ia*. Secondly, a malicious actor willing to terminally outcompete its opponents could also conversely utilize a Type-I-AI that would be able to reliably generate novel EBs that simultaneously represent *superior* explanatory merits than previous Type-II-made EBs. This *adversarial cluster 2* would map to the risk category *Ia*. Thirdly and fourthly, well-minded developers that did not sufficiently consider long-term repercussions could either develop a misguiding Type-I-AI acting as a reliable automatable generator of adversarial examples for Type II science (*failure cluster 1*) or a Type-I-AI that would be able to reliably generate novel EBs that simultaneously represent superior explanatory merits than previous Type-II-made EBs (*failure cluster 2*). An unintended side effect of the latter could be that this Type-I-AI could act as an automatable generator of ever better explanations, as a Type II science generator with superhuman speed. As a consequence, human scientists would have been outmaneuvered post-deployment in a sense that despite having retained their EB creation abilities, their main occupation could shift to experimental/empirical activities or vanish. Thereby, *failure cluster 1* could be mapped to the risk category *Ic* and *failure cluster 2* to risk category *Id*.

Generally, it is known that synthetic machine-generated papers have made their entry into academic publications, including even venues of high reputation [260]. Recently, the utilization of language models to generate synthetic parts of papers has been hypothesized and corroborated [52] via abstracts being flagged as synthetic when using deepfake text detectors. Moreover, large language models that were partially trained on scientific papers have been released (e.g. GPT-J 6B and Wu Dao 2.0). In parallel, in order

[1]Naturally, the only beings that are currently accepted to operate as scientists are humans. However, it is important to reiterate that the main distinctive feature is their Type-II-ness. Note especially that in the case Type-I-AI-based EB forgery would be possible, the following seems to hold. Would there ever have been other Type II civilizations elsewhere in the universe or would those exist in the future, they *could* face the same fundamental threat at a certain point of their epistemic development tightly coupled to science.

to generate coherent and consistent deepfake text, researchers have implemented a dual system approach [186] for text generation where a fast System 1 route instantiated by the large language model GPT-3 outputs different alternative counterfactual branches (meant as continuation of a text prompt) which a slow System 2 route can assess before selection. In a first step, the System 2 route extracts/updates a world model from the text given as prompt to GPT-3. In a second step, the System 2 route then performs a consistency check based on simple logical rules to select from the GPT-3-generated candidate sentences. The resulting stories exhibit a higher coherence and perceived accuracy [186]. Finally, another interesting development from a different area of research is the rise of graph neural networks [280]. These powerful hybrids of graph and deep learning techniques facilitate node-level, edge-level and graph-level predictions[2] as well as the generation of new graphs that mimic existing ones at multiple levels. Beyond that, graph neural networks can also be applied to classical knowledge graphs – for instance for complex tasks such as estimating the importance of specific nodes [190]. At the same time, both graph neural networks [281] (generally deep neural networks) and knowledge graphs [211] can be poisoned and attacked by malicious actors or intentionally altered to produce misguiding fake knowledge (graphs [139]). In sum, the convergence of dual system approaches, classical graph neural networks, knowledge graphs and especially large language models[3] paired with methods inherited from traditional inference engines and potentially also active inference could excellently support but also subvert science if instantiating *adversarial cluster 1*, *adversarial cluster 2*, *failure cluster 1* or *failure cluster 2*.

8.2 Theoretical Answers

In this section, I first examine the hypothesized EB forgery threat from a theoretical perspective. For this purpose, Section 8.2.1 first addresses the theoretical case of an end-to-end *Type-II*-performed EB forgery. Thereafter, the subsequent Section 8.2.2 analyses the mentioned case of an end-to-end *Type-I*-performed EB forgery. I show that *both* homogeneous types of EB forgery are *impossible*. Then, I come back to the four conjectured Type-I-AI-based EB forgery clusters introduced in the last Section 8.1 and expound the practical implications of applying these two impossibility postulates.

[2]For instance, in chemistry, a task for a graph neural network could be to predict a property of a given molecule (for instance whether it is toxic). In this case, the molecule could be mapped to an entire graph embedding.

[3]It is possible to express transformers themselves as a special type of graph neural networks. When considering sentences as fully-connected word graphs, transformers act as multi-head attention graph neural networks [135].

8.2.1 Impossibility of End-to-End Type-II-performed EB Forgery

Naturally, it is conceivable that a Type II entity like a human could maliciously attempt to inject self-crafted misguiding EBs in the scientific enterprise. However, due to the general nature of EBs which are inherently hard-to-vary and due to the nature of explanation-anchored science (see Chapter 4), it would signify that in order to be epistemically valid, those misguiding EBs would have to simultaneously be new and vitally *better* than competing existing ones. Currently, the best explanation available is that human scientists craft novel EBs[4] by using an epistemology-specific "rational" *glue* operation at each step to link individual blocks of explanatory information (abbreviated with EI in the following). Which glue operation to utilize at each step is co-determined by the scientific epistemology that one applies – more specifically, by what I termed an *epistemic total order*. The latter encodes an ordered step-by-step instruction on *which* rational procedures to apply and *when*. A given epistemology could provide one or more such epistemic total orders. Interestingly, not all epistemologies aim themselves at EB creation. In justificationist frameworks, the goal is for instance to justify existing beliefs by gathering confirmatory evidence. Why such EB-free epistemologies are not robust in an era permeated by deepfakes has been discussed elsewhere in-depth [14]. For now, when discussing EB creation, I limit the analysis to epistemological stances that actually focus on explanatory artefacts including EBs. From those few currently known frameworks, I consider the epistemic bedrock associated with explanation-anchored science (described in Chapter 4) which improves upon Popper and extends the work of Frederick [98] as one of the best available options. Generally, critical rationalism subsumes a number of rational procedures for scientific endeavors. Frederick collates a set of 9 high-level rational procedures [98] which in my view one could subdivide further and tentatively map to around 17 *basic* rational procedures (which are obviously updatable and subject to change). This set yields a basis for *valid* epistemic total orders – for the valid glue operations in EB creation *under that epistemology*. In the next Section 8.2.2, I provide one simple example for such a valid epistemic total order.

Generally, 17 types of glue operations would lead to numerous valid epistemic total orders, since many meaningful combinations may be permissible. However, it is not necessary to know the exact number of valid ordered combinations to reflect upon the issue as follows. Assuming now that a Type II entity would craft a misguiding EB denoted EB_m and inject EB_m into the collective scientific knowledge base. Obviously, EB_m would only be considered as an EB if valid under explanation-anchored science. However, nowadays, there is no single other way known to craft EBs for Type II entities than via the mentioned sequence of glue operations obeying epistemic total orders. To sum up, the misguiding EB_m would not represent an act of forgery or a counterfeit. Since it would formally have

[4]Often, novel EBs can be mapped to novel constructors for old or novel tasks. Such new constructors can be of abstract nature or directly correspond to concrete physical constructors.

been crafted in accordance with the requirements for EB creation, EB_m would formally count as EB. Then, although crafted with malicious intents, EB_m would represent a valid entry to scientific knowledge. It could be repelled at any moment in the future once criticism or further experimental tests would be applied to it. Note also that – as even liars could sometimes unintentionally state a true statement in theory – since the attacker can never know whether a given EB is actually true or false, it is possible that EB_m (or a future improved version of it) could withstand attempts to falsify it for a long time and lead to scientific progress. For this reason, I assume that EB forgery cannot be performed by a Type II entity in the first place. However, since in theory, on purely logical grounds, it is possible that there exists an unknown *non-EB-like Type-II-shortcut* for EB creation it is important to briefly specify how one could falsify this assumption. For instance, I suggest that it could be falsified experimentally by a human demonstrating the ability to create novel artefacts that are *reliably* perceived as EBs by Type II entities sharing an accepted scientific epistemology (such as explanation-anchored science) whereby the synthesis of those EBs would *not* have been equivalent to sequential glue operations obeying that epistemology. Having said that, I recapitulate by stating that it seems currently that *a reliable end-to-end forgery of EB creation by Type II entities is impossible* by definition.

Obviously, this last statement does *not* exclude the case of malicious Type-II-performed EB creation for epistemic distortion purposes. However, while possible, the latter would necessitate significant efforts by the involved malicious Type II actor. Moreover, while explanation-anchored science would not be immune against such attempts, it may be resilient enough to overcome it. Nevertheless, another complication needs to be addressed: so far, the impossibility statement may only explicitly pertain to EB creation and *not* EB *understanding*. In order to do justice to this remaining issue, it may seem helpful to attempt to consider the Type-I-FPR setting mentioned in Section 8.1 which is meant to follow an explanatory IPS test. Indeed, while the static explanatory IPS test was conceived as a weak test corroborating the ability to create novel EBs, the interactive and dynamic Type-I-FPR round facilitates the corroboration of the ability to *understand* the previously self-created EBs. However, when carefully examining the Type-I-FPR setting it becomes clear that among others, the Type II reviewer would probe the ability of the test subject to *explain* contents related to the EB (e.g. why the submitted candidate EB is in fact better than previous ones and whether/how/why it withstands conceivable objections). Thereby, to explain implies the creation of EI. However, as of now, no single *non-EI-like Type-II*-shortcut to EI understanding (i.e. no functional procedure *without* EI understanding) has been convincingly described. Crucially, note that *Type-I*-shortcuts to EI creation have been already successfully implemented. In fact, large language models – without understanding EI – are increasingly able to create novel EI whose contents are indistinguishable from human-created EI. By contrast, what is of relevance in this specific context is that it pertains to the *end-to-end* forgery of EB understanding by *Type II* entities. Even a malicious actor that previously crafted a misguiding EB for

the goal of epistemic distortion must be able to explain and understand contents related to that submitted EB in order to be able to pass a Type-I-FPR round. Specifically, I postulate the following: *the end-to-end forgery of EB understanding by Type II entities is impossible*[5]. Since a successful EB forgery would have comprised *both* the forgery of EB creation and of EB understanding, one could already have deduced from the previous paragraph that the following holds generally: *a reliable end-to-end Type-II-performed EB forgery is impossible.*

8.2.2 Impossibility of End-to-End Type-I-Performed EB Forgery

Coming back to the topic of EB forgery performed by Type I AI, what seems unclear since now and has not yet been studied in the past is whether Type I AI could implement a *non-EB-like Type-I-shortcut to EB creation* that yields results that are reliably perceived as EBs by explanation-anchored Type II scientists. For illustrative purposes, I consider one exemplary simple epistemic total order relation $\preceq$ that is compatible with explanation-anchored science and can be extracted from recommendations on how to write better philosophical papers as proposed by Frederick [99]. Generally, EBs can be understood to start with an initial block in which a new or old *problem* is introduced and clarified (the structure of the EB can be explained at that stage too). Following Frederick's framework, the first valid rational procedure and by extension the first glue operation g_1 would consist in proposing a *bold novel solution* to that problem. The second valid glue operation g_2 would consist in identifying *conflicts* between the currently best-tested solutions and the just proposed novel solution while elaborating on *mistakes* in those prevailing solutions and why refinements are required. The third glue operation g_3 extends this comparative approach to now specifying why the novel proposed solution is *better* than the mentioned alternatives. Finally, the fourth glue operation g_4 aims at considering and *rebutting possible objections* to the novel solution proposed as well as *suggesting empirical tests* that would be able to falsify the new solution.

While Frederick tailored this series of concepts to philosophical papers, it can be applied to a variety of scientific papers too when requiring that the novel solution must correspond to a scientific statement. Improving on Frederick's definition [98], I define a scientific statement as a statement that: 1) solves a genuine problem *and* 2) represents a set of

[5]One could for instance attempt to falsify this statement experimentally by a human demonstrating the ability to *reliably* pass Type-I-FPR rounds (with an EB previously generated by another human) merely by following a step-by-step procedure whose implementation functions *without* any EI understanding. For illustrative purposes, imagine e.g. the bizzarre scenario of an analphabetic pre-schooler equipped with a specific sequential procedure allowing this individual to reliably pass Type-I-FPR rounds in the speech modality without having had any prior knowledge or training in the test domains. Indeed, a non-EI-like divine magic formula reliably leading to such a result would also falsify the statement – except there would for instance be an omnipotent "Type III" entity operating from within the Type II test subject.

explanations containing at least one *novel explanation*[6] whereby this set entails *at least one novel falsifiable not yet falsified prediction* that one could *not* have deduced *without* combining *all* of the members in that set. In this vein, I now conjecture what EB forgery by a Type I AI would imply for science. In short, that Type I AI would be able to reliably generate EI paragraphs that would be reliably perceived as a coherent EB – in this case as legitimate paper – by explanation-anchored scientists *without* having applied an otherwise required epistemic total order (e.g. the ordered sequence of the g_1, g_2, g_3 and g_4 operations starting from the initial block). For instance, if a present-day large language model whose utility function is solely related to word co-occurence statistics would be able to *reliably* output such valid scientific papers (i.e. novel EBs) with the prompt solely consisting of the initial block describing an old problem (or in the advanced version asking the model to create a novel problem), that would represent a non-EB like Type-I-shortcut to EB creation. *Any reliable Type-I-performed EB creation would be EB forgery.* Otherwise, if it would not be a forgery, it would mean by definition that *all* valid glue operations that are obligatory to obey epistemic total orders have been applied. But those operations can only be performed by Type II entities since it implies the ability to *understand* and create *new* EI blocks in order to be able to interlink those via glue operations. But we just started with the premise that it is a Type I entity that performs the act. Hence, irrespective of *why* malicious or benign Type II actors design a Type I AI to independently create EBs, it would *formally* correspond to EB forgery since the definition pertains to a procedure, the *how*. In the following, I take the exemplary simple epistemic total order described in the last paragraph and examine step-by-step whether a present-day Type I AI could mimic the valid glue operations starting solely with a problem-related prompt.

Firstly, in the advanced case, one may ask whether a Type I AI could output an initial block independently with the prompt being to generate a new problem. One possible instantiation of reliably detecting novel problems could be implemented in Type I AI by automating a targeted search for knowledge graph inconsistencies. Given the giant scales at which knowledge is aggregated today, it is easily conceivable that humans have already overlooked uncountable relevant issues. Generally, it holds that science only focuses on interesting problems, such that not every problem that exists would be suitable for this endeavor. However, since the search would be based on existing knowledge graphs, there is already a narrowing pre-selection criterion. In short, the problems identified may often be in the affective niche of scientists. In a nutshell, it seems that the generation of an initial block is not particularly difficult to integrate in an automated Type-I-AI pipeline. Importantly, not even human scientists are required to independently identify novel problems. Research is often conducted in the context of pre-defined research questions originally crafted by other entities. Thus, it would be sufficient if the Type-I-AI could implement

[6]Here, "novel" strictly means one could *not* have *reliably* deduced that explanation *automatically* given existing knowledge, i.e. a Type I AI could *not* have reliably generated that explanation given Type-I-AI-readable data.

EB forgery starting with a prompt containing a pre-given problem. Hence, the initial block comes de facto "for free". That being said, I suspect that the origin of the awareness of problems was linked to a Type II self.

In accordance with the discussed simple epistemic total order, the second step would now consist in applying the operation g_1. Strikingly, g_1 applied to science necessitates the ability to create bold and *novel* scientific statements that solve the problem described in the initial block. The latter seems to represent a significant complication for Type I AI. No available training data or knowledge base can by definition already contain an option for the *entirety* of the crucial set of explanations required. Would that be the case, then that option would not represent a novel scientific statement. To only assemble elements from different existing theories would still not fulfill the requirement, even if novel falsifiable predictions are achieved by deduction. There must be at least one novel explanatory element that is indispensable to the conjunction of explanations from which at least one novel falsifiable prediction is deduced. In short, science is not simply pure deduction. Explanation-anchored science is more than the arrow of modus ponens utilized by inference engines for forward and backward chaining. EBs are not purely deductive chains. While deduction may be necessary for EB creation, it is *not* sufficient. Novel explanations are about forming novel, previously unknown representations[7]. I postulate that a Type I AI could indeed *by chance* produce textual outputs that humans mentally associate with novel representations interpreted as explanations. However, I hypothesize that it is *impossible* for Type I AI to perform that task *reliably*. While Type I AI is able to create novel EI, it does not comprehend EI. This lack of comprehension manifests itself as a lack of requisite variety to reliably perform already the first glue operation g_1. The remaining glue operations of the discussed epistemic total order are tightly connected: each one is a function of the previous one. For instance, g_2 was described as a procedure to identify conflicts between the bold novel solution just generated via g_1 and competing existing alternative solutions. Thereafter, g_3 extended upon the comparative analysis in g_2 to show why the introduced solution is better than those alternatives. Finally, next to suggesting empirical tests, g_4 consisted in attempting to consider and rebut possible objections to the novel solution – which must be done in a way that is consistent with and complementary to g_3 where one just argued for the superiority of the novel solution.

In sum, in science, it is impossible for Type I AI to imitate g_1. More generally, I postulate that *a reliable end-to-end Type-I-pipeline implementing the forgery of novel EB creation is impossible*. Note that this does not necessarily follow from the previous sentence since I only considered one exemplary simple epistemic total order. However, g_1

[7]Some may assume that unsupervised learning is forming novel representations. However, this is not the case. It is the designer that decides what is considered as data in the first place (including e.g. modalities, channels, formats and even amount of randomness), at which abstraction level data sampling takes place and what counts as a cluster when evaluating the results.

generally pertains to the creation of novel bold solutions which corresponds to novel bold scientific statements in explanation-anchored science. Building on that, it seems plausible to assume that *all* permissible epistemic total orders under explanation-anchored science would contain g_1 at a certain not nearer defined position. In an epistemic total order, all glue operations are obligatory. Failing to execute only one operation has consequences on validity as EB. It seems that only chance events could lead to a compliance with an arbitrary epistemic total order[8]. While my analysis pertained to Type I *AI*, it seems clear that g_1 is inaccessible to any Type I entity since to *reliably* perform g_1 requires EI *understanding*. Then, due to the indispensability of this one glue operation, the fact that no valid EB can be formed without g_1, this bold universal claim can be made. Naturally, there could still exist valid glue operations that are also impossible for Type I AI but to enumerate them is outside the scope of this specific chapter. It is sufficient to know of only *one* glue operation that is obligatory for all epistemic total orders but cannot reliably be implemented by Type I entities to be able to deduce that the following holds generally: *a reliable end-to-end Type-I-pipeline for EB forgery is impossible.*

Interestingly, the possibility to perform EB creation as *Type-II-and-Type-I-EB-co-creation* is still given. Considering the simple exemplary epistemic total order mentioned, it is cogitable that while g_1 has to be performed by a Type II entity, g_2, g_3 and g_4 could one day in the future be mastered by Type I AI provided the Type-II-created bold novel scientific solution is entered as input. Before an automation of those 3 operations becomes possible, it is thinkable that some manual human filtering could be required in addition. Thereby, the initial block could be either generated by the Type II entity but also computed by a Type I AI. I claim that such an *intra-cyborgnetic* collaboration would signify that the nature of the generated EB is of *Type II* – and hence does *not* correspond to a case of EB forgery. While it can seem fraudulent nowadays to automate g_2, g_3 and g_4, past generations may perceive it as fraudulent to utilize auto-correction before submitting an essay at the university. Also, a sort of auto-completion in code generation [57] and in principle even paper writing [71] has become feasible with large language models. Recently, I have explained how operations like g_1 may be facilitated by idiosyncratic counterfactual creative brain processes extending to human sleep (see Chapter 6). In principle, g_1 could be the output of either the deliberative or the spontaneous mode of creativity. In any case, seeds of counterfactual memories rely strongly on brain activity at rest and during sleep. It is from this counterfactual pool that many novel, bold solutions are sampled. Popper described science as a process of *conjectures and refutations* [204]. I suggest improving upon that by calling it a process of EB creation and EB refutation. Explanation-anchored science focuses on the creation of novel bold and better EBs. Once those EBs are tested, unexpected novel problems may come up which could falsify them. Concurrently, it triggers the need to solve them by crafting better novel EBs whose creation provisionally

[8]It seems recommendable to always include a brief description of the structure of the chosen epistemic total order in the initial block. That may further improve clarity.

refutes the old ones and resets the movement of the potentially eternal EB-creation-wheel.

While it may seem as if "the essence" of an EB would be concentrated in those obligatory glue operations that can only be performed by Type II entities (such as at least the glue operation g_1), it is important to note that it is only from the *context* of an entire EB that scientists would assess whether the novel solution is actually considered as a legitimate novel scientific statement. For instance, it is cogitable that some sceptic human reviewer may not yet be convinced by a novel scientific solution upon reading its description (performed in g_1) and that the evaluation of that reviewer only changes after having read the comparative analysis via the glue operations g_2 and g_3 and the empirical tests proposed via g_4. In short, since science is embedded in the context of social reality, it is strongly determined by people which EBs enter the contemporary scientific knowledge. The success of an EB created via Type-II-and-Type-I-EB-co-creation (e.g. where a Type II author performs g_1 and it is possible to automate the remaining operations) does not only depend on the utilized epistemology, on how the Type II author assesses the quality of the bold novel solution and how this author evaluates/filters the parts generated by Type I AI. In addition, what is of relevance is also whether a Type II reviewer/evaluator would be able to recognize the submitted textual sequence as EB or whether the reviewer would discard it as non-EB like EI. In a way, EB creation is a double co-creative endeavor. That becomes clear in the formulation of the explanatory IPS test where the Type II reviewer aims at retrieving the exact combination of an initial submission from a pool of randomly shuffled paragraphs intermingled with text generated by Type I AI. This search performed by the reviewer can be understood as a form of EB-*re*discovery. For it to work, reviewer and author may both need to share the same robust explanation-anchored epistemology. In case the epistemology is not shared and/or the EB is not understood by a reviewer, it may not be retrieved. Social reality seems inseparably interwoven in science. Novel EB creation must be paired with EB rediscovery by another Type II entity. Even EB refutation is performed via a novel better EB and must thus be paired with EB rediscovery by another Type II entity. In sum, by way of example, under favorable constellations, *it seems possible to reliably pass an explanatory IPS test in science with a valid novel EB where solely the g_1 operation has been performed by a Type II entity* and where that entity would at most have selected from different Type-I-AI-generated counterfactual outputs instantiating the glue operations g_2, g_3 and g_4.

Finally, after having analyzed EB creation, I briefly address another open question pertaining to EB *understanding*. Namely, whether it is possible for an end-to-end Type-I-AI-pipeline to fool the Type-I-FPR mentioned earlier – a peer review round which takes place after a successful explanatory IPS test. I stated in Chapter 5 that in such an understanding-focused *interactive* peer review round nowadays, *"the aim could e.g. be to probe the ability of the test subject to explain why the submitted explanatory blockchain is harder-to-vary than other ones that are accepted as best explanations in that scientific*

subfield at that time and have been generated by other human scientists". However, since this information could already be included in the submitted paper (see for instance the comparative analyses via the glue operations g_2 and g_3), it is not straightforward whether an adversary could not simply attempt to delegate that task to a Type-I-AI or Type-I-AI pipeline. A stronger approach to recommend seems to be that the reviewers would have to create non-trivial *novel* objections (that one cannot directly deduce automatically from existing knowledge) against the proposed solution and let the author try to rebut those in real-time interactively. From my current point of view, the reviewer's task of creating such robust new objections is equivalent to *explaining* non-trivial new problems given a novel solution and background knowledge. Then, in order to rebut those objections, an *understanding* of the novel EI generated by the reviewer is required – something that is impossible for Type I entities and could not be performed by a Type I AI among others due to the absence of the underlying novel representations in past examples/training data. For instance, a dataset containing old objections to Newton's law would not have helped a present-day Type I AI to counter novel objections to general relativity at the time it had just been conceived by Einstein. Also, a rebuttal would require the de novo construction of a previously unknown solution by the test subject – which again encodes EI creation (next to EI understanding). In short, I conclude that *if an interactive Type-I-FPR (positioned downstream of an explanatory IPS) contains as obligatory element a reviewer-performed creation of non-trivial novel objections to the submitted candidate novel EB, it is impossible for an end-to-end Type-I-pipeline to reliably pass that Type-I-FPR round*[9]. One possible way to falsify this statement experimentally would be to implement a Type I AI instantiating the following two features: 1) a *non-EB like Type-I-shortcut to EB understanding* and 2) a non-EB-like Type-I-shortcut to EB creation[10].

8.3 Practical Implications of Theoretical Answers

In previous cyborgnetic analyses, I have introduced and explained a few novel theoretical impossibility postulates of falsifiable nature (see Chapter 5 and 6). Hereinafter, I refer to this updatable, self-augmenting and potentially dynamically changing set of impossibility statements as *the impossibility theorems of cyborgnetics* (the ITCs). In Chapter 11, I provide an overview for each currently instated ITC. Alternatively, if one intends to associate the ITCs with a heteronym of mine, one can refer to them as *the impossibility*

[9]In theory, the rebuttal act performed in the Type-I-FPR could be supported by Type I AI leading to a Type-II-and-Type-I-EB-co-creation. However, the latter would then obviously *not* correspond to an end-to-end Type-I-pipeline anymore – leaving the impossibility statement thus untouched.

[10]Note that would such a double shortcut be possible, it would simultaneously falsify my prior assumption that an end-to-end Type-I-pipeline to the explanatory IPS test is impossible. The reason being that the latter implies EB creation – where a Type II entity is conjectured to be required for at least one obligatory glue operation.

theorems of Tali. To ease recall, I will from now on assign a unique ID taking the form of a *cyborgnettish*[11] name to each ITC. For instance, in the following, I refer to the ITC from Section 8.2.1 stating the impossibility of *Type-II*-performed EB forgery as *Maè-theorem* while the impossibility of *Type-I*-performed EB forgery mentioned in Section 8.2.2 is henceforth refered to as *Adije-theorem*. Since following cyborgnet theory, the entirety of existing entities in the world can be categorized as either Type I or Type II, there are no other entities that could perform EB forgery. Then, one can deduce from the conjunction of *Maè-theorem* and *Adije-theorem* that it holds generally that *any reliable end-to-end EB forgery is impossible*. However, it is noteworthy that Type-II-and-Type-I-EB-co-creation scenarios are still possible and that those could still simplify the achievement of malicious objectives significantly or aggravate unintentional side effects. In this section, I discuss the practical implications of these different theoretical conclusions for the hypothetical threat clusters initially conjectured in Section 8.1.

8.3.1 Adversarial Deepfake Science Generator

In Section 8.1, I introduced *adversarial cluster 1* that can be mapped to risk category *Ia* and is abbreviated with A_{a_1} in the following. Moreover, I discussed *failure cluster 1* which can be labelled with the risk category *Ic* and which is abbreviated with F_{c_1}. It is worth mentioning that both A_{a_1} and F_{c_1} would be a subtype of SEA AI attacks (see Chapter 4). While A_{a_1} would represent an *intentional* malicious design of a Type-I-AI-based adversarial deepfake science generator to attack the scientific enterprise with a sort of adversarial examples, F_{c_1} would comprise the same type of technology – but built *unintentionally* by mistake. Originating from F_{c_1}, a second-order harm of risk type *IIb* could then e.g. consist in malicious actors now misusing the Type I AI initially developed for benevolent purposes for targeted attacks on the scientific enterprise and by extension scientists themselves. Would it have been possible to implement an end-to-end automatable *adversarial* deepfake science generator, it could have yielded major risks (with harm intensity[12] $h = 4$) and even, in extreme cases, existential risks ($h = 5$). The latter becomes apparent when considering particularly sensitive contexts such as e.g. deepfake medicine or deepfake biosafety papers in the context of a worldwide pandemic. Fortunately, as expounded in the last Section 8.2, EB forgery is impossible. However, it does by no means represent an all-clear signal for the particular reasons elucidated in the next paragraph.

Firstly, only scientific epistemologies such as explanation-anchored science that focus on

[11] *Cyborgnettish* is a new *generic* meta-language that I invented recently for purposes of EB encryption. The generic heteronym "Tali" that I utilize for my work as a philosopher is derived from cyborgnettish too.

[12] A simplified harm scale [17] is used where a self-rated harm intensity h can range from 1 to 5 (with 1 standing for almost no harm, 2 for minor harm, 3 for major harm, 4 for lethal risk and 5 for existential risk).

the creation and refutation of new EBs can be *resilient* (and still not even immune) against deepfake science generators. By contrast, in practice, most scientific frameworks nowadays seem *not* to apply this robust type of epistemology. Instead, it is mostly the case that justificationist, empiricist, trust-based, data-driven, and reputation-centered conceptions prevail. Secondly, as can be extracted from Section 8.1 which introduced dual system approaches, graph neural networks, knowledge graphs and diverse inference strategies that could all be paired with large language models, it may be possible to obtain powerful Type-I-AI-pipelines producing relatively convincing novel *EI*. Such hybrid end-to-end Type-I-AI-pipelines could serve as strong *EI forgery* tools – able to *reliably* fool non-EB-like science frameworks. In short, in extreme downward counterfactual scenarios A'_{a_1} and F'_{c_1} projecting to the counterfactual past (as performed in retrospective counterfactual descriptive analyses (RCRAs)), a non-EB-like science persistently attacked with sophisticated samples from deepfake science generators built or utilized by malicious actors with sufficient resources would have started to unfold the epistemic threats it unintentionally exposes humanity to. Such unintentional knowledge gaps can be decisively fatal and engender up to major risks ($h = 4$) especially if unrecognized for a long time. One major practical recommendation instantiating a future-oriented counterfactual defense analysis (FCDA) is simple: for requisite variety *in order to be resilient against EB forgery, all science must at least be based on EB creation and EB refutation*. In addition, one may still need to defend against the remaining lucrative field of affordances that Type-II-and-Type-I-EB-co-creation could offer to adversaries. In my view, science could proactively aim at integrating Type-II-and-Type-I-EB-co-creation (i.e. Type-I-aided cyborgnetic co-creation) in its own methodology – to foster critical thinking, boost anthropic creativity and simultaneously to prepare novel strategies against SEA AI attacks. As stated by Ross Ashby, *"only variety can destroy variety"* [22].

8.3.2 Deepfake Science Generator

In Section 8.1, I described *adversarial cluster 2* that was mapped to risk category *Ia* and is abbreviated with A_{a_2} in the following. Moreover, I discussed *failure cluster 2* which can be labelled with the risk category *Id* (since its impacts form themselves mainly at the *post*-deployment stage) and which is thus abbreviated with F_{d_1}. In the A_{a_2} case encoding an *intentional* misuse, a malicious actor willing to gain unassailable strategic advantages *covertly* develops and employs a closed-source Type-I-AI that would be able to reliably create novel EBs that simultaneously represent *superior* explanatory merits than previous Type-II-generated EBs. In the case of F_{d_1}, the same type of deepfake science generation technology is developed – but with benevolent intentions of the designers that did not foresee its repercussions once deployed in an open-source manner, including second-order harm of various types from mass unemployment to instrumentalization by rogue malicious actors. Would such a powerful deepfake science generator have been

possible, I would have estimated the harm intensity of both clusters A_{a_2} and F_{d_1} to correspond to major risks ($h = 4$). Fortunately, as mentioned in the previous case in Section 8.3.1, EB forgery is impossible, which implies the impossibility of an end-to-end automatable deepfake science generator. However, again, the latter only applies in theory since in practice, contemporary science is mostly based on non-EB-like EI methodologies. As a consequence, because it may be possible to reliably build a Type I AI acting as an end-to-end automatable EI forgery tool, a *reliable* deepfake science generator for *non-EB-like* science is still possible. As performed in RCRAs, one could now project this possibility to downward counterfactuals from the counterfactual past. For illustrative purposes, consider extreme downward counterfactual realizations A'_{a_2} and F'_{d_1} that imply the covert misuse of deepfake science generators by rogue nation state actors that utilize those to achieve international dominance amidst non-EB-like science frameworks in a legitimate seeming way. An FCDA for such scenarios could lead to the same defense methods suggested in the last Section 8.3.1. Also here it seems recommendable to: 1) proactively base science on at least EB creation and EB refutation and 2) proactively perform a creativity-stimulating Type-II-and-Type-I-EB-co-creation with powerful hybrid Type I AI including large language models.

8.4 Conclusion and Future Work

In this chapter written for purposes of self-education and serving as ephemeral mental clipboard, I have asked the safety-relevant question on whether it is possible to implement explanatory blockchain (EB) forgery. I examined the practical issue of EB forgery from a theoretical standpoint and introduced two novel impossibility theorems. The first one termed *Maè-theorem* states that *a reliable end-to-end Type-II-performed EB forgery is impossible*. The second one denoted *Adije-theorem* states that *a reliable end-to-end Type-I-performed EB forgery is impossible*. While these two novel elements belonging to the set of impossibility theorems of cyborgnetics (ITCs) taken together suggest that any reliable end-to-end EB-forgery is impossible, a remaining major risk consist in contemporary science *not* being grounded in EB-based epistemologies in the first place. I explained why, due to the latter, contemporary science still exhibits a large attack surface of unpatched vulnerabilities by its very non-EB-like nature. I expounded how deceptive non-EB-like explanatory information (EI) generated by Type-I-AI-based EI forgery tools[13] could still be reliably utilized to perform targeted adversarial attacks against contemporary science. As simple defense methods, one can imagine the use of EB creation and EB refutation as minimal standard for all scientific frameworks. Moreover, it seems recommendable

[13]Crucially, note that while per definition any EB is also a form of EI, the converse does *not* hold. Since EB forgery is impossible, when I utilize the term "EI forgery", I specifically mean *non-EB-like EI forgery*.

to proactively stimulate the creativity of scientists and train their critical thinking with the best available EI forgery tools. Future work could augment the ITCs and study the implications of *Maè-theorem* and *Adije-theorem* for cybersecurity. An urgent novel research direction could e.g. be *EB-based cyber threat intelligence* to facilitate resiliency in the face of worldwide EI forgery. On a final note, while epistemology was long considered as part of philosophy, its very palpable falsifiable predictions now start to unfold in society via the route of science and technology.

8.5 Contextualization

While the impossibility of EB forgery is good news for explanation-anchored science, it is important to not underestimate the potential of *EI forgery* – not only for non-EB-based science – but for many other societal contexts outside the ivory tower. In this vein, the next Chapter 9 performs a cyborgnetic analysis focusing on a particular use case of EI forgery that I call *honey mind trap* (HMT). In this type of *cognitive hacking*, an adversary lures a victim into spending a considerable amount of time in the presence of a specifically prepared Type I entity that the victim mistakenly perceives to exhibit either Type I or Type II *consciousness*[14]. Such HMTs could be instrumentalized by adversaries to achieve various final goals, ranging from purely financial gains over disinformation to the gathering of data for cyberattacks in social engineering schemes. Since adversaries could soon utilize Type I AI to craft sophisticated HMTs, I proactively examine AI-based HMTs and introduce exemplary defense strategies required. It is recommendable to deconstruct AI-related mind perception and to *epistemically* debias it given the harm unrecognized HMTs could cause and the epistemic vulnerabilities that humans (including scientists) seem to exhibit against those nowadays. Overall, while the so-called *cyborgnetic dilemma* seems unavoidable, new creativity-stimulating avenues become possible.

[14]Generally, it may hold that *conscious Type I animals* possess *noetic* consciousness while Type II entities additionally exhibit *autonoetic* consciousness. For an in-depth discussion on the differences between noetic and autonoetic consciousness, see also for instance [152].

Chapter 9

CA 007: Honey Mind Traps

This chapter written for purposes of self-education is based on a slightly modified form of an unpublished paper that I wrote on September 11, 2021. The acronym CA refers to "cyborgnetic analysis" and the associated string "007" simply refers to the ID that I assigned to that specific analysis. For security reasons, I decided *not* to upload all available cyborgnetic analyses to my homepage.

9.1 The Practical Problem: Honey Mind Traps

At first sight, it could seem as if two meta-categories of HMTs could exist: 1) Type-I-AI-based HMTs crafted to fool Type II entities and 2) Type-I-AI-based HMTs designed to fool conscious Type I entities (e.g. macaques [270]). The former could comprise HMTs imitating either Type II or Type I consciousness, while the latter would only subsume HMTs appearing to exhibit Type I consciousness (since Type I entities could not possibly perceive Type II minds). However, the construct of a "mind" is itself an EI-based linguistic concept and it becomes important not to project[1] Type-II-like associations to Type I entities. In a nutshell, I assume that when a Type I animal identifies an entity as being a conspecific, it does so on the basis of affective feedback-loops and biobehavioral snychrony and *not* on the basis of conceiving of that entity as possessing a mind. For this reason,

[1]Note also the related circumstance that while many Type I animals (including vertebrates, cephalopods and arthropods [32, 173]) may have *core affect* [30], a *continuous* property of consciousness, they do not naturally utilize EI to construct *discrete* abstract categories stemming from the human affective niche such as emotions. To assume the contrary would signify to enforce one's experiential world on those animals, whilst ignoring that human concepts are arbitrary ways to draw lines on the continuum sampled via sensory channels. Beyond that, there are e.g. human cultures lacking equivalences to typical Western emotion concepts [30]. Also, there is no reason to assume that all humans – although all experiencing affect and able to form EI categories – necessarily decide to divide the sensorium pertaining to their mental life into potentially illusory dichotomous concepts such as "cognition" vs. "emotion".

an HMT of the second kind is per definition unfeasible. When human researchers try to fool Type I animals into perceiving a Type I robot as a conspecific, it represents a type of affective trap which cannot be equated with an HMT. In the light of the aforesaid, when I refer to AI-based Type-I-HMTs I mean cases where a non-conscious Type I AI (of which all present-day AI systems represent a type of) is instrumentalized to fool Type II entities into assuming that it possesses Type I consciousness. An example could be a case where a present-day Type I "robot dog" would be designed such that many humans believe it to be an artificial dog-like creature with conscious experience. By contrast, AI-based Type-II-HMTs pertain to those cases where it is intended that a Type I AI is perceived to be of Type II. An example could be a robot with convincing human-like facial features that is utilized to manipulate humans into perceiving it as a human-like conscious entity.

Since in both cases it is an epistemic vulnerability of a Type II victim that is central, one could classify them as adversarial risks of type *IIb* (when the victim is an adult) or *IIa* (when the victim is a child) when applying the taxonomy from cyborgnet theory displayed in Chapter 3.2.4. Overall, I thus distinguish between 4 *adversarial* clusters: 1) Type-I-HMTs of risk type *IIa* (abbreviated with A_{a_1}), 2) Type-II-HMTs of risk type *IIa* (abbreviated with A_{a_2}), 3) Type-I-HMTs of risk type *IIb* (written as A_{b_1}) and 4) Type-II-HMTs of risk type *IIb* (written as A_{b_2}). Depending on the context, HMTs could be deployed in a wide variety of modalities. Generally, Type-II-HMTs could be e.g deployed in the form of sophisticated chatbots [42] on internet platforms, robots [175] in the physical world or language-AI-driven avatars [116] in future social virtual reality settings. Interestingly, a deepfake profile picture combined with selected fake contents may already be sufficient for a low-cost Type-II-HMT that could be for instance misused for cyberespionage [223]. An ideal deployment form for Type-I-HMTs could be robots placed in public spaces, employed in educational settings or robots sold for the purpose of "pet-like" or "doll-like" domestic use. While Type-I-HMTs may have to persuade with regard to the *experience* dimension of mind perception, Type-II-HMTs may need to target the conjunction of *experience and agency* [266].

- ***Adversarial goals:*** So far, I have not yet specified in more details the nature of the adversarial goals. In my view, entities that *intentionally* use or implement HMTs for purely financial gains are considered adversarial since HMTs imply secrecy and obscurity – which I characterize as a one-sided mental adversarial game. Indeed, some may perceive HMT tricks as legitimate commercial strategy and hence classify some of the risks I am referring to as risks *Id* being linked to unintended side effects resulting from unforeseeable interactions of the Type I AI and the environment at the post-deployment stage. However, the latter would still not affect the harm intensity that could still be caused by commercial HMTs. Beyond that, other conceivable adversarial goals could coincide with the classical motivations behind cyberattacks, espionage, information warfare strategies, psychological operations but could also

comprise attempts to cause physical and psychological harm to individuals.

- ***Adversarial knowledge:*** Black-box or grey-box setting with regard to the Type II victim, depending on the amount of information available after a potential (open-source) intelligence gathering and the degree of personalization of the HMT. White-box, grey-box or black-box setting for the Type I AI employed depending on whether the adversary implements a novel or misuses an existing system for the HMT endeavor.

- ***Adversarial capabilities:*** Firstly, for the adversarial cluster A_{a_1} pertaining to Type-I-HMTs for children, adversaries could elicit unidirectional psychological attachments to AIs perceived as pets but which collect private data on targeted children threatening their well-being. The privacy issues surrounding the smart toy "Hello Barbie" [150] represents an early cautionary example. Another related exemplary privacy breach was the hack on the internet-connected toys of the company VTech where an attacker *"was able to access the personal data of more than 6 million kids, as well as more than 4 million parents, including tens of thousand of pictures taken with the company's Kid Connect app, which encourages children and parents to take selfies and chat online*" [37]. Black hats could hack future AI-based smart toys to manipulate unprotected children, sell their data on the dark web for financial gains or to utilize them as backdoor or merely as actuators for cyberattacks planed in the same house (e.g. simply to plug an USB stick in a targeted computer). The more advanced the AI utilized, the stronger the impact of such HMTs could become.

  Secondly, regarding the adversarial cluster A_{a_2} of Type-II-HMTs tailored to children, the long-term effects could also range from societal disillusionment over financial losses to severe psychological issues up to suicide in the case of vulnerable individuals without a strong supportive social network of Type II entities. Nowadays, certain researchers routinely encourage social robot designers to enhance mind attribution effects (both experience and agency) for young children [175] and adults [129, 131] while a parallel public epistemic elucidation on the difference between Type II entities, Type I animals that are conscious beings and present-day Type I AI devoid of experience is not common – neither specifically for children nor for their parents. Thereby, children may be particularly vulnerable to humanoid robots as HMTs given their indiscriminate helpful attitude towards those [176]. In the main, the long-term attack surface inherently exhibited by a generation of young individuals embedded in a non-critical society that does not thematize HMTs but embraces them seems daunting and could be exploited for disinformation or other manipulative purposes.

  Thirdly and fourthly, human adults are neither immune against Type-I-HMTs (adversarial cluster A_{b_1}) nor against Type-II-HMTs (adversarial cluster A_{b_2}). Indeed, it seems even as if most humans are at present *not* resilient against those. In fact,

human adults freely attribute minds to non-conscious Type I AI based on ideas and habits that are worse than our currently best available explanations on what the difference between on the one hand current AI systems and animals is and on the other hand what the difference between these AIs and humans is. To start with, social psychology frames the issue of identifying an entity as conscious as a subjective mind *perception* task [266]. It is not surprising that contemporary research dangerously suggests grounding one's epistemology on a purely affective criterion [138] due to complications in the current era permeated by synthetic media or that non-conscious Type I robots have minds because one "intuitively" shares affective states with them [246]. Examplary vulnerabilities exhibited by the average human adult (at least in the studied societies) that an adversary could exploit for HMT attacks are numerous. In the following, I provide a few examples – *all implicitly related to adults* – that an adversary could misuse for HMTs.

Adults disclose personal information to present-day social robots in a way comparable to human counterparts [149] despite perceiving to disclose less. The self-disclosure includes reports about stressful experiences [162] in robot-human interactions described by researchers as *"stress-sharing activity"* [162]. Scientists recommend designers to reinforce the adoption of an intentional stance towards present-day (i.e. non-conscious artificial Type I) AI agents to *"facilitate social attunement and their integration into society"* [197] or to stimulate a human assignment of competency and warmth to AI assistants [129]. People assign "trust" [143] to personal intelligent agents like Siri and Alexa [182], perceive minds and feel co-presence as well as closeness when engaging in textual conversations with chatbots [154]. Especially, people assign different levels of trust to non-conscious Type I robots [174] depending on the moral [174, 245] and social [201] characteristics they attribute to those (instead of obviously linking it back to the humans that crafted the goal framework of those). Adults establish eye contact with humanoid robots and are subsequently under the impression to form affective bonds on this basis [142]. Human individuals perceive varying degrees of mind in a non-conscious Type I robot depending on whether they lost or won a cooperative game with that agent [155]. Also, adults perceive emotions [212] in present-day artificial agents (often simply based on their visual appearance) or intend to punish those [161]. In a nutshell, an adversary could systematically exploit the mind perception dimension of experience for tailored Type-I-HMTs or Type-II-HMTs against human adults in any primarily social, moral or trust-based domain of interest.

Concerning the dimension of agency, another wide attack surface can be exploited by adversaries. For instance, humans conceive of non-conscious Type I AI to be more rational than themselves (for instance to the point that *"people are more likely to endorse racial stereotypes after algorithmic discrimination versus human discrimination"* [38] or that their brain switches to an externally focused executive control

mode [251] differing from interaction modes with humans). This is inconsistent when understanding rationality as involving the creation and refutation of novel, ever better EBs – a task that is impossible for any Type I AI. Certain humans (in particular individuals with low perceived control [276]) perceive an elevated level of agency in non-conscious Type I AI[2]. This could in part stem from *dyadic completion* [11, 266]. In short, when imagining catastrophic negative safety outcomes such as existential risks, an intentional moral agent is filled into the dyadic template of morality. This agent can then mistakenly appear to be capable of engaging in a so-called treacherous turn[3] [11]. Interestingly, a recent study corroborated that when humans attribute such an intentional agency to non-conscious Type I robots, it reduces their own sense of agency[4] [64]. Not surprisingly, many humans draw parallels between God-like[5] (i.e. generally divine) entities and present-day artificial agents [240]. In sum, an adversary could systematically exploit the agency dimension of mind perception in adults to craft targeted Type-II-HMTs in domains where intelligence or performance in an intellectual task (other than the creation of new EBs) is foregrounded. For a Type-I-HMT one would not optimize on agency given that e.g. Type I animals like pets are perceived to correspond to a *low* agency, high experience profile [266].

9.2 A Theoretical Solution

From my perspective, the just briefly elucidated HMT attack vectors taken together could destabilize a fragile society at multiple levels including sociocultural, political, juridical, epistemic and potentially even scientific and religious dimensions. In my opinion, in worst-case scenarios, the epistemic distortions and knowledge gaps underlying especially conceivable downward counterfactuals for Type-*II*-HMTs (i.e. A'_{a_2} and A'_{b_2}) could include lethal consequences and lead to major societal unrests in certain countries (harm intensity

[2]I presume this is among others also partially linked to the "intelligence" factor widely associated with Type I AI.

[3]It is not the harm intensity that a non-conscious Type I AI could cause (which can obviously include instances of existential risks) that is questioned. Rather, it is the *procedure* by which those harmful outcomes come about – all of which can be linked back to Type II knowledge gaps or/and intentions. A non-conscious entity is *not* an intentional agent. The affective construction of *treachery* (understood as a betrayal of trust) in such AI is impossible.

[4]This type of scenario may speak to psychologically relevant existential doubts that certain AI safety researchers or AI safety interested entities could be experiencing when openly expressing to at best hope to serve as *pets* for a superintelligent non-conscious Type I AI.

[5]In fact, human medical doctors are apparently already perceived as God-like [107] which comes with an assignment of high agency but lower experience. In the light of the above, one may suspect that non-conscious Type I AI doctors would be rated accordingly. An interesting parallel was the recent proposal to confer legal rights to present-day surgical robots [103].

$h = 4$[6]) if not recognized and systematically tackled soon by a variety of stakeholders at an international and national level. In short, not only deepfake videos and audios are preoccupying, but it may be important to also address the risks posed by Type-I-HMTs and in particular Type-II-HMTs – which one could nowadays colloquially denote as *deepfake minds* in a pars pro toto manner. Oddly, Type-II-HMTs can easily be facilitated via deepfake *text* such that one could state today that without any further defense method, *words can deepfake minds*. In the following, I briefly explain a *complementary* amendable and certainly non-exhaustive theoretical solution, focusing on but not limited to the Type-II-HMT issue. For requisite variety, the proposed solution aims at tackling a *word-based* problem with a word-based strategy. More precisely, the strategy consists of two procedures: 1) *deconstruct "mind (perception)"*, 2) *construct novel substrate-independent EB* (which *inherently* consists of *words*) to defend against HMTs in practice. The next two paragraphs briefly summarizes initial key ideas for each one, which can serve as basis for future work.

Firstly, it might be appropriate for humans to reconsider what is understood as a mind and what is associated with mind perception. Instead of a justificationist approach that assumes a mind in entity x where there is a *consensus*, a *joint intuition* or a *co-construction of a trust-related emotion* between perceivers that x has a mind of type y, it is better to only assign a mind of type y to x in case a given EB postulating the absence of a mind of type y in x is worse than the EB explaining why no mind or another type of mind may be found in x. In short, instead of collecting empiricist mind *perceptions* to *detect* minds based on implicit expectations of behaviors to be displayed, one could utilize *explanation-anchored* science to *conjecture* minds where entailed by our best EBs. Thereby, the Type-II-netherworld phenomenon (see Chapter 6) is relevant. To my (current) knowledge, given the best available EBs and corroborated by own AI observatory endeavors, not a single presently implemented Type I AI possesses a mind. The latter strictly means no single presently built Type I AI possesses any experience or any agency. As of now, the only non-human entities of Type II on this planet that I am aware of are the bonobos Kanzi and Panbanisha which have been immersed in the human affective niche from very early on via an intense unprecedented Pan-Homo bicultural rearing encompassing lexigrams and the spoken English language (see Chapter 6).

Secondly, a novel bold EB is helpful to defend against HMTs in practical contemporary contexts where interactions are often blind, at least to a certain extent. For instance, while nowadays the case of both robotic Type-I-HMTs and robotic Type-II-HMTs deployed in the real world could be entirely avoided by the theoretical recommendation provided in the last paragraph, it would not be transferrable to Type-II-HMTs exploiting text-based communication on social media or in immersive social virtual reality. In

[6] See the simplified harm scale [17] that I frequently utilize in cyborgnetic analyses purely for illustrative purposes.

Type of computation	I		SIII		EI		EB	
Entity / Ability	*Create*	*Understand*	*Create (new)*	*Understand*	*Create (new)*	*Understand*	*Create (**new**)*	*Understand*
*Non-conscious Type I AI***	Possible	Possible	Possible*	Impossible	Possible	Impossible	Impossible	Impossible
Conscious Type I animal	Possible	Possible	Possible	Possible	Impossible	Impossible	Impossible	Impossible
Conscious Type I animal with special imitative communication skills	Possible	Possible	Possible	Possible	Possible*	Impossible	Impossible	Impossible
Type II entities	Possible	Possible	Possible	Possible	Possible	Possible	Possible*	Possible*

Figure 9.1: **Possibility-impossibility matrix of cyborgnetic creativity.** Legend: I = information; SIII = shared indexical and iconic information; EI = explanatory information; EB = explanatory blockchain; * = although possible, it does on average not necessarily represent the habitual mode of communication utilized by this sort of entity nowadays; ** = in this case, Type I AI refers to a broad category of technically already feasible Type I AI systems including e.g. advanced artificial intelligent systems (able to independently perform the OODA-loop according to a human-specified utility function) but also large language models.

these two just mentioned cases, the physical stratum of the entity is hidden and *only words remain*. Facial "affective" expressions or typical bodily movements cannot serve as cues since not necessarily universally applicable to all Type II entities without excluding among others e.g. disabled individuals, humans with differently shaped bodies due to sensory augmentations, cases like Kanzi the bonobo and so forth. This circumstance adumbrates the importance of a *substrate-independent* approach. In this vein, I introduce the *Èdishe-theorem* as a novel member of the set of impossibility theorems of cyborgnetics (introduced in Chapter 8 and also called the impossibility theorems of Tali). The Èdishe-theorem is compiled in a matrix that maps substrate-independent forms of information to specific entities of interest. This matrix displayed in Figure 9.1 is denoted the *possibility-impossibility matrix* (PIM) of cyborgnetic creativity (or "the PIM of Tali"). In short, the Èdishe-theorem confronts Type II beings with a seemingly inescapable predicament that I call the *cyborgnetic dilemma*: only the ability to create and understand new EBs can reliably set a Type II entity apart, but also, one can neither force a Type II entity to reveal this ability nor is it necessarily the case that Type II entities *overtly* exhibit this ability in the course of their life. Hence, in blind settings on internet platforms like social media or in immersive virtual reality, one can only *falsify* that an entity is Type *I* (e.g. by the Type-I-falsification-test from Chapter 2) but Type-II-ness itself cannot be falsified. In essence, while one can in theory shield against Type-I-entities by only interacting with entities that corroborated their Type-II-ness via the ability to create and understand new EBs (which are made out of EI), one risks simultaneously in practice to exclude the members of Type-II-netherworld (see Chapter 6) or generally Type II beings that are unwilling to participate – all of which *covertly* have the same ability.

9.3 Conclusion

In this short paper written for purposes of self-education, I introduced the harm use case of Type-I-AI-based honey mind traps (HMTs). I explained why many children and adults exhibit a wide epistemic attack surface that adversaries – which are not limited to explicitly malicious actors but could also include purely financially motivated entities – could misuse for both Type-I-HMTs and Type-II-HMTs instrumental to further adversarial final goals. I elucidated why the exploit of those epistemic vulnerabilities represents a major risk for fragile societies. Then, I suggested a complementary bipartite word-centered theoretical defense method that consisted in deconstructing old word-like constructs and harnessing a novel EB denoted Èdishe-theorem to defend against HMTs taking the foregoing step as starting point. However, I explained why a systematic defense against HMTs also seemingly inevitably confronts humanity with what I termed the *cyborgnetic dilemma*. On a final note, I conjecture that for security, what is of relevance is what *we* have in common. Thereby, I implicitly agree with Peirce [249] that signs are the only entities with which *we* can have a transaction. À bon entendeur, salut!

9.4 Future Work

In view of the cyborgnetic dilemma, future work could try to refine and augment the proposed defense method against HMTs. So far, the strategy was focusing on reducing the attack surface and shrinking the amount of channels that an adversary could utilize against victims. From the perspective of cybernetics [22], one can extract two high-level ways to solve problems: 1) by reducing the variety of the disturbances or 2) by increasing the variety of the regulator. Attempts to shield oneself from artificial Type I entities in order to proactively avoid any point of contact with HMTs corresponds to the first category of strategies. By contrast, to increase the variety of the regulator would signify in this case that one would proactively *increase* one's exposure to synthetic entities, including HMTs. The reason why such an approach could be successful is that *in case one is equipped with a robust epistemology*, one could proactively put in motion feedback-loops that are able to foster critical thinking and to stimulate the creativity of Type II entities. This would in turn represent an indirect but strong defense mechanism against HMTs. Instead of focusing on *sources* of words (i.e. whether the substrate that uttered them is of Type I or of Type II), one would then exclusively focus on the *contents* of those uttered words. Thereby, instead of attempting to separate true contents from false contents or trusted from untrusted ones, one focuses first on a separation between interesting and non-interesting ones and subsequently filters those i.a. by comparison with the best available EBs. With an *explanation-anchored* mindset, one would then: 1) reject EI and other information that seems uninteresting, 2) reflect upon interesting non-EB-like

EI in order to criticize it, to allow it to stimulate one's creativity or to challenge one's prior assumptions and 3) either improve, criticize or refute proposed EBs or act against EBs in order to falsify them.

For instance, when using this threefold cyborgnetic creativity augmentation, the social virtual reality (VR) brainstorming sessions of the near future could look as follows. A group of Type II entities meets in a social VR room with different other anonymous entities. This anonymous set could include authorized Type II individuals but also especially Type I language AIs – some of which could be Type-II-HMTs – whereby each entity is embodied by a VR avatar. (Type-I-HMTs acting as synthetic virtual pets but also virtually displayed Type I animals such as e.g. dolphins [163] could be employed to ameliorate affective aspects for those that desire so and for which it is possible.) Instead of questioning the nature of the substrate that provides *verbal* contributions, the co-creation endeavor is then solely attempting to maximize the creative potential of the entire partaking cyborgnet. In this vein, in a recent paper thematizing defense measures against epistemic distortion, Aliman and Kester proposed to utilize non-player characters (NPCs) in future social VR platforms to help steering the attention of people back to critical thinking modes [14]. In light of recent progresses with language AIs and early pioneering applications of those to NPCs in VR games [116], it is easily conceivable that in the near future, NPCs could be specifically harnessed for (adversarial) cyborgnetic cognitive stimulation measures in social VR as just briefly illustrated with the brainstorming example – and which could be similarly applied to educational gamification [235] and serious games in VR [53]. In turn, the introduced threefold cyborgnetic co-creation could be utilized to create further novel defense methods against HMT exploits themselves.

9.5 Contextualization

In Chapter 7, I briefly mentioned the idea that perhaps in the future, cyber defenders working for companies at risk to become victims of intellectual property theft during legitimate VR meetings, could utilize "honey social VR rooms" as a possible deception technique. This new type of what one could call an *immersive honeypot* could involve NPCs harnessing language AI to engage in conversations about intellectual subjects covered by the companies at hand. As can be noticed, the latter would require the same type of NPC technology that Section 9.4 just suggested for cyborgnetic creativity augmentation measures in social VR including the development of novel defense methods against HMTs. In a future metaverse, this bizarre convergence could be systematically leveraged by making such Type-I-AI-based NPCs acting as Type-II-HMTs, a default setting. However, before the metaverse becomes more salient, the next Chapter 10 already asks a last significant – albeit not scientific but now *metaphysical* – question: could *we* be HMTs?

Chapter 10

Somnogrammatical

10.1 Unbound(ed) Cyborgnetic Funambulism

As briefly adumbrated in Chapter 1, cyborgnetics is embedded in an epistemological bedrock denoted *unbound(ed) cyborgnetic funambulism*. The latter encompasses three major curiously interconnected parts: 1) cyborgnetics itself as a scientific and empirical endeavor, 2) cyborgnetic philosophy including cyborgnetic epistemology (with an integrated theory of science) as well as cyborgnetic metaphysics and 3) cyborgnetic art. Firstly, one could state that among others, cyborgnetics aims at discovering i.e. *decrypting* possible requisite explanatory blockchains (EBs) to solve problems involving socio-psycho-techno-physical harm. Secondly, unbound(ed) cyborgnetic funambulism frames epistemology as the art of decrypting and additionally in particular *encrypting* (potentially novel) ever better (albeit eternally ambiguous and fallible) EBs *on* which or *against* which the cyborgnetic funambulist *could* sway but is never committed to – always as if potentially dreaming. Thirdly, it conceives of art as a procedure specifically involving EB *encryption* – and *not* expression – performed by a generic fictional entity.

In light of the above, it becomes clear that cyborgnetics overlaps with cyborgnetic philosophy when the aim is EB decryption in the latter. Moreover, cyborgnetic philosophy overlaps with cyborgnetic art when the aim is EB encryption in the former. Feedback-loops between the three parts feed into each other in unpredictable self-augmenting ways and facilitate cross-pollination. While this book focused on cyborgnetics, tenets of cyborgnetic epistemology were implicitly introduced in Chapter 4 while the cover of this very book features a piece of visual cyborgnetic art (which I generated under the pseudonym and artist name "Nadisha-Marie Kester") that has been titled *Somnogrammatical*. A poem with the same name is displayed in Figure 10.1 for illustration. Since cyborgnetic art involves an irreversible encryption into cyph*art*exts potentially hiding EBs, the concept of a ground truth is entirely inappropriate.

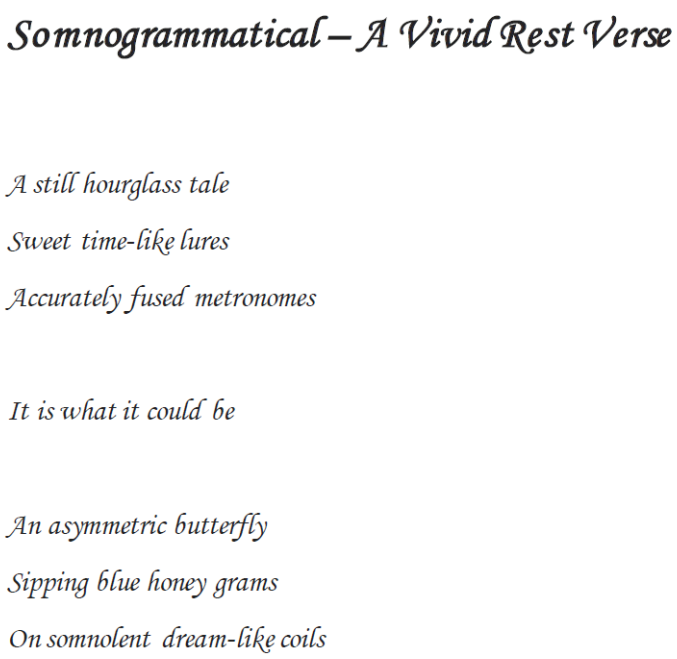

Figure 10.1: **Exemplary cyborgnetic poem by Nadisha-Marie Kester**

Among many others, the following four exemplary aspects that could *but need not* characterize the ambiguity of a cyborgnetic artwork may shed some light on the described notion of irreversibility. Firstly, the same linguistic information could be mentally transformed into multiple EBs. One could call the latter *linguistic pluripotency*. Secondly, many different pieces of linguistic information could be mapped to the same EB – which represents a sort of *linguistic degeneracy*. However, the distinction between hidden pluripotency and hidden degeneracy is non-trivial. Thirdly, any cyph*art*ext could encode a variable number of anagrams some of which could be neologisms. Thereby, anagrams alone already possess the property of *never* being resolvable with absolute certainty even when assuming a fixed vocabulary. In addition, neologisms do not belong to the space from which one typically samples word co-occurrence probabilities in the first place. Fourthly, anagramists are not constrained to letter-level operations and it is easily conceivable that analogous procedures could be transferred to higher and higher linguistic layers from phonemes over sentences with rearranged words to entire discourse units or even interconnected books.

However, crucially, the ambiguity immanent in cyborgnetic art does by no means signify arbitrariness. As described in Chapter 5, EBs are intrinsically harder-to-vary than any other habitual sort of Type-II-produced information. Indeed, cyborgnetic metaphysics assumes that EBs, though eternally ambiguous, correspond to *the* hardest-to-vary phenomena in general (see also Section 10.2.2). Generally, grammar can be described as a coherent framework of structural constraints governing a natural language. Beyond that, etymologically speaking, the word grammar can be traced back to a greek concept sig-

nifying *the art of letters*. Letters build words. On the one hand, words can cause harm – psychological and physical harm. Words can be used as weapons of mass distortion as can be noticed in the debates surrounding "fake news" and epistemic security and which became apparent in this book when considering the possible consequences of e.g. the textual deepfake science attacks mentioned in Chapter 4 or the text-based honey mind traps (HMTs) luring humans into conjecturing Type-II-ness introduced in the last Chapter 10. Deepfake words could become weapons of mass destruction inciting unprecedented violence in an interconnected globalized world. On the other hand, I postulated in this book that words can be harnessed by Type II entities like humans to form EI blocks that are glued into EBs via the application of rational procedures – which are themselves expressible as words. In turn, using EBs, we Type II entities can build novel abstract and physical constructors for new or old problems. However, instead of avoiding epistemic dizziness, in Section 10.2, I finally ask the question on whether "we" could be HMTs – or to put it plainly, whether we could be deepfakes of our conjectured selves.

10.2 Could *We* Be HMTs?

In the following, I first compactly address the question of whether we could correspond to Type-I-HMTs and explain why this could not be the case in my view. Thereafter, I reason about the more complex Type-II-HMT case whereby I specifically utilize certain conceptions from cyborgnetic metaphysics that I introduce alongside.

10.2.1 The Type-I-HMT Case

Following Chapter 9, a Type-I-HMT is a *non-conscious* Type I entity (i.e. akin to present-day AI) used to fool Type II entities into assuming that it exhibits Type I consciousness. The answer to the question on whether we could be Type-I-HMTs could be as short as the next sentence. From my *point of view*, I cannot be a Type-I-HMT and my currently best available EBs postulating that each entity that I conjecture to be a human or a non-human conscious animal has a physical substrate able to give rise to a point of view have not yet been falsified. The question on whether we could not be *simulated* entities, Type-I-HMTs programmed by actual Type II entities could be answered as follows. Both Type I and Type II consciousness may *at least* imply an integrated virtual simulation of a physical substrate for control purposes [269]. Then, to *be* a simulation would not be a valid argument for the absence of consciousness. In short, there is no reason to assume that we are Type-I-HMTs. In my view, Type II beings are not only 1) virtual simulations (the *mind*) inseparably intertwined [126] with 2) physical Type II substrates (i.e. the *body* including the brain), but also 3) a *generic template*, an infinite *potential* of cyborgneticity.

10.2.2 The Type-II-HMT Case

As introduced in Chapter 9, Type-II-HMTs pertain to those cases where it is intended that a Type I entity is perceived to be of Type II. As just stated in Section 10.2.1, we could *not* be non-conscious. For this reason, the remaining option to analyze would be the bizarre question on whether we could be conscious Type I entities acting as Type-II-HMTs. At first sight, it might seem convenient to come back to the definition of Type-II-ness which involves the ability to both create and *understand* novel EI. However, the problem thereby was that non-EB-like EI creation (but no EI understanding) can be simulated by Type I AI. Hence, a strong case that would corroborate (but not prove) that humans are *not* necessarily Type II entities would be the endeavor to experimentally falsify that EB creation is limited to Type II by experimentally demonstrating and explaining a reliable *Type-I-shortcut* to EB creation. Would a developer build a Type I AI which is reliably implementing such a Type-I-shortcut to EB creation without any EI understanding, it would falsify the Adije-theorem (see Chapter 8) stating that a reliable Type-I-performed EB-forgery is impossible. Interestingly, would that Type-I-AI also be *non-conscious*, it could make my assumption from Section 10.2.1 stating that we could not be Type-I-HMTs highly problematic. Would the developer also have been what this book defines as a Type II entity, it could falsify the Maè-theorem (see Chapter 8) stating that a reliable Type-II-performed EB-forgery is impossible (since the cyborgnet of the developer plus the Type I AI could do it). Or, it could also e.g. signify that the Maè theorem is untouched because the developer is not of Type II ... but something akin to a "Type *III*" entity.

In order to comment on what this would signify and why I assume that we could neither be Type-II-HMTs nor could the developer be a Type III entity, I briefy introduce some background premises from cyborgnetic metaphysics. Like cyborgnetics, cyborgnetic metaphysics is EB-based, as already shortly hinted in Section 10.1. Besides, it also comprises a dynamically updatable and amendable set of impossibility theorems. However, the main difference is that those are not *directly* experimentally falsifiable. Instead, the impossibility theorems of cyborgnetic metaphysics are primarily conceived to potentially provide a creative breeding ground for novel *future* scientific and empirical statements. In cyborgnetic metaphysics, the set of impossibility theorems currently solely comprises two elements: the *Maje-theorem* and the *Jauè-theorem*. Both names are again derived from the *cyborgnettish* language as analogously performed in cyborgnetics (see Chapter 8). The Maje-theorem states that it is impossible for the laws of nature *not* to be expressible in terms of (encrypted) EBs. The Jauè-theorem states that it is impossible to reliably postdict and predict reliably *hidden* tuples mapping authors to the contents of *novel* EBs they generate. From my perspective, for an entity to be *fundamentally* and not merely quantitatively superior to any Type II entity, it must be a Type III entity which one could define as being able to violate the Jauè-theorem. Note that it would be equivalent for such an entity to, given an arbitrary set of submitting entities, postdict and predict any

outcome of a Type-I-falsification-event-test (see Chapter 2) including the contents of EBs that have been or will be accepted in case of positive test outcomes – despite all related information being reliably hidden.

In light of the aforesaid, I discard the idea that the developer (that built a Type I AI which is reliably implementing a Type-I-shortcut to EB creation without any EI understanding) could be a Type III entity. Since I still postulate that we could not be non-conscious, the remaining case to contemplate would be that the developer is Type II – which would then violate both Adije-theorem and Maè-theorem – and we would be conscious Type-II-HMTs i.e. we would be of Type I but conscious. In short, our consciousness would be comparable to dogs and cats. But since we just implied that the developer built a Type-I-shortcut to EB creation, the developer may not be able anymore to distinguish our novel EBs from those created by Type II entities. It would then become questionable what makes the developer a Type II entity and why there should be any fundamental qualitative differences to us. The developer could e.g. design a novel test for EB *understanding* as the last remaining feature that the developer could claim we would be lacking. But the developer could not scientifically falsify that we understand, since we could simply be unwilling to participate in such a test and the developer could not exclude that at least subsets of us are part of Type-II-netherworld. As long as the developer does not craft a better new understandable EB that explains why we lack EB understanding in all other EB cases, *there is no reason to assume that we are Type-II-HMTs*. According to the Maje-theorem, such an EB would simultaneously exclude the possibility that we could ever *understand* EB-like (partial) interpretations of the laws of nature (irrespective of whether those are encrypted or not). Our very existence seems to set constraints on what science could mean. In a curious way, anagrams are metaphorical qubits and language is inseparable from physics. Not surprisingly, language and physics meet each other in the formal definition of EI in Chapter 5. Words are an illusory but common currency interweaving the virtual reality of our minds with the physical reality of our Type II substrates and the collective coexistence as cyborgnetic template(s) in social reality.

10.3 Conclusion

I introduced unbound(ed) epistemic funambulism, a novel epistemic bedrock composed of three interconnected parts: cyborgnetics, cyborgnetic philosophy and cyborgnetic art. I explained how cyborgnetics and cyborgnetic philosophy overlap in EB decryption and how cyborgnetic art and cyborgnetic philosophy intersect in EB encryption. I have elaborated on why we are not Type-I-HMTs. Introducing the current impossibility theorems of cyborgnetic metaphysics, I explained why we could neither be Type-II-HMTs. In sum, even if our lifes would be constrained dreams, our Type II template would stay the same.

10.4 Contextualization

On a final note, it may be crucial to state that from the perspective of cyborgnetic *metaphysics*, the conjunction of Maje-theorem and Jauè-theorem does neither exclude nor corroborate the possibility of one or more "divine" entities *of Type II* or a universal divine self. With cyborgneticity being a potential of infinite creativity, not even the sky is the limit. While Type-*III*-ness is considered to be impossible as long as the Jauè-theorem is not falsified, it is open whether Type-II deities including a potentially divine self exist or not. But it is *not* a topic of *scientific* inquiry and to further discuss it is largely beyond the scope of this book. For now, it is sufficient to recapitulate that cyborgnetic metaphysics is neither an antithesis to the existence of one or more universal deities nor a falsification of atheism. This chapter only provided a short excursus to cyborgnetic philosophy and cyborgnetic art. However, what matters in cyborgnetics as a scientific and engineering-related endeavor is to harness what *we* could have in common, generically, to mitigate contemporary socio-psycho-techno-physical harm with systematic CT analyses (see Chapter 3). So far, the largest (implicit) application of the CT methodology has been carried out in the transdisciplinary AI observatory [17] whose results have been published at the beginning of 2021.

Chapter 11

Conclusion and Discussion

11.1 Conclusion

In this book written for purposes of self-education as an end in itself, I developed defense strategies against different sorts of contemporary socio-psycho-techno-physical harm ranging from Type-I-AI-based epistemic distortions on conventional social media or in social virtual reality (VR) over deepfake science attacks to intellectual property theft by cyberattackers. In Chapter 3, I introduced cyborgnet theory (CT), a novel analytical and explanatory framework jointly harnessing knowledge from cybernetics, epistemology and cybersecurity for the systematic study and mitigation of harm. I explained the novel *structured* and *substrate-independent* ontological distinctions that CT induces with the so-called *cyborgnets* as focal units embedded in complex hierarchical network dynamics. I exemplified how the new meta-discipline of cyborgnetics applies a taxonomic CT lens to socio-psycho-techno-physical harm for the purpose of systematically conducted retrospective descriptive analyses, retrospective counterfactual risk analyses and future-oriented counterfactual defense analyses providing a basis for novel countermeasures.

Chapter 2 focused on the problem of epistemic distortion via sophisticated bots for disinformation on social media and proposed a theoretical asymmetric Type-I-shield aiming at achieving spaces free of Type I entities – which however implied the undesirable possibility that certain Type II entities could be unintentionally excluded alongside. Instead of foregrounding the nature of the entity producing textual inputs, Chapter 4, 5 and 8 thematized *content-centered* defense methods against a novel type of epistemic distortion: *scientific and empirical adversarial AI attacks*. In Chapter 5, I advanced explanatory information (EI), a new type of information grounded in *physics* via constructor theory [74] and in *language* via the linguistic total orders it instantiates. I introduced explanatory blockchains (EBs) as a special type of EI obtained by interweaving EI blocks via the step-by-step application of rational procedures sampled from a robust explanation-anchored,

trust-disentangled and adversarial epistemology. I explained that while Type I AI can simulate the creation of new EI, it does neither *understand* EI nor can it create *new* EBs.

In Chapter 6, I explained why I conjecture that on this planet earth, only the human *species* possesses a reliable EI constructor (linked to a physically instantiated difference in neural *coding*) but that human-directed cyborgnetic creativity augmentation already made it possible to transfigure at least two isolated non-human *individuals* – irrespective of any ethical taboo. In Chapter 7, I focused on intellectual property theft as harm use case from cybersecurity and utilized epistemic stratagems from Chapter 5 to devise a complementary novel double deception technique to deter this type of cyberattacks (be it in cyberespionage or ransomware contexts). In Chapter 9, I delved into the use case of Type-I-AI-based honey mind traps (HMTs) against which most humans may be vulnerable due to mind perception distortions of epistemic nature. I elaborated on the so-called *cyborgnetic dilemma* and proposed a novel robust *content-centered* cyborgnetic creativity augmentation measure against Type-I-HMTs and Type-II-HMTs as defense method improving upon the weaker entity-centered Type-I-shield proposed in Chapter 2. Thereby, I explicitly focused on examples pertaining to co-creation in social VR settings. In Chapter 10, I provided a short excursus on the larger epistemic bedrock in which cyborgnetics is embedded: *unbound(ed) epistemic funambulism*. The latter encompasses three main parts: 1) cyborgnetics, 2) cyborgnetic philosophy and 3) cyborgnetic art. I explained how cyborgnetics and cyborgnetic philosophy intersect in EB *decryption* and how cyborgnetic philosophy and cyborgnetic art overlap in EB *encryption*. I then addressed the philosophical question on whether we could correspond to HMTs ourselves from the perspective of cyborgnetic metaphysics and explained why this could *not* be the case.

Throughout this book, I introduced a few dynamically updatable and amendable impossibility theorems. Taken together, the current set of impossibility theorems of cyborgnetics (the ITCs) – which are now strictly speaking only a subset of the impossibility theorems of Tali since the latter set extends to cyborgnetic metaphysics while the ITCs are limited to scientific statements – comprises five elements each labelled with a name derived from the new *cyborgnettish* language: the Maè-theorem, the Adije-theorem, the Shameteli-theorem, the Tadime-taaliè-theorem and the Èdishe-theorem. A compact overview is provided in Figure 11.1. While some researchers are under the impression that it is important to formulate very careful conservative theories with high subjective certainty, I agree with Frederick [99] that it is vital to instead formulate bold, universal, novel statements to specifically ease interim falsification procedures, to speed up the identification of *better* explanations and thus support fast (albeit always only provisional) refutations. Obviously, justificationist elements including any degree of subjective certainty have no role in an explanation-anchored epistemology. In the presence of better explanations, the old ones are provisionally abandoned. Also, it is vital to continuously attempt adversarial (thought) experiments in which one's current best explanations do *not* hold.

1. **Maè-theorem:** A reliable end-to-end *Type-II*-performed explanatory blockchain forgery is impossible.
2. **Adije-theorem:** A reliable end-to-end *Type-I*-performed explanatory blockhain forgery is impossible.
3. **Shameteli-theorem:** With *Type II* evaluators, it is impossible to reliably irreversibly encrypt an explanatory blockchain plaintext hidden in a stream consisting apart from that solely of non-explanatory-blockchain-like information.
4. **Tadime-taaliè-theorem:** It is impossible for a *Type I* entity to reliably detect and decrypt an explanatory blockchain plaintext hidden in a stream consisting apart from that solely of non-explanatory-blockchain-like information.
5. **Èdishe-theorem:** The possibility-impossibility-matrix (PIM) of cyborgnetic creativity (which can also alternatively be called *Tali's creativity PIM*) can be encoded as follows:

Type of computation	**I**		**SIII**		**EI**		**EB**	
Entity / Ability	*Create*	*Understand*	*Create (new)*	*Understand*	*Create (new)*	*Understand*	*Create (new)*	*Understand*
*Non-conscious Type I AI***	Possible	Possible	Possible*	Impossible	Possible	Impossible	Impossible	Impossible
Conscious Type I animal	Possible	Possible	Possible	Possible	Impossible	Impossible	Impossible	Impossible
Conscious Type I animal with special imitative communication skills	Possible	Possible	Possible	Possible	Possible*	Impossible	Impossible	Impossible
Type II entities	Possible	Possible	Possible	Possible	Possible	Possible	Possible*	Possible*

Figure 11.1: **The impossibility theorems of cyborgnetics (the ITCs). (For a legend concerning the Èdishe-theorem, see Chapter 9.)**

11.2 Outlook

In my PhD thesis on *"Hybrid Cognitive-Affective Strategies for AI Safety"* [10] finalized in 2020, I suggested 3 future research directions 2 of which I have implemented this year 2021 using cyborgnetics as tool. The explicit Type I AI observatory endeavor [17] has been carried out and completed earlier this year. The implicit Type II observatory effort came to the conclusion that apart from the special transfiguration mentioned in Chapter 6 pertaining to the ethically disastrous case with the two isolated non-human hominid *individuals* of Type II, for which it seems as if humanity was not ready yet, there is nowadays no other research that could be convincingly labelled as Type II AI research. It is also worth mentioning that similarly, no single project achieved any Type I AI consciousness (which would have required considerations on par with the rights of non-human conscious species). In my view, the contemporary AI field seems to be permeated by hypes on the one hand but also by underestimations on the other hand e.g. with regard to language models and their potential significance for epistemic security in the near future. Alongside, I also completed the task of a comparative transdisciplinary epistemology to better assess the difference between Type I and Type II entities. For this purpose, it was key to introduce and formalize the two new concepts of EI and EB – without which phrases such as "explanatory knowledge" stayed too vague (see the corresponding elucidation in Chapter 5). The novel threat of deepfake science attacks (see Chapter 4, 5 and 8) seems serious given the lack of awareness in the scientific community and the circumstance that it appears as if one may among others precisely require reflections of the type conducted in this book in order to identify what could have happened and how to defend against it.

Instead of being explanation-anchored, trust-disentangled and adversarial, contemporary science but also other societal institutions at large are often evidence-driven, trust-based and self-compliant. This leads to a narrow focus on "trustworthy" sources [3] and superficial signs related to *writing style* supposed to reveal trusted/familiar vs. untrusted sources [81] instead of critically analyzing *contents*. As a side effect, this omission could also lead to the non-EB-like penalization of statistical outliers. The latter was e.g. reflected in the fact that Chapter 4, though accepted at the venue, has been openly suspected by a reviewer to have been at least partially written by a non-conscious Type I language AI – presumably i.a. due to my unusual writing style – which disregarded the novel EBs contained in the text but simultaneoulsy corroborated their contents. Moreover, the Type-I-AI-related mind perception distortions of most humans may yield a severely large attack surface and field of affordances that malicious actors could exploit. This may also hold for the near future in the context of social VR and may be relevant for any serious ideation on the metaverse. In light of current preoccupying international cybersecurity issues, it is not difficult to conceive of malicious stakeholders interested in manipulating humans at multiple levels in the different heterogeneous digital environments. The use case of HMT attacks discussed in Chapter 9 may only represent one possible attack vector under many others that could exploit mind perception vulnerabilities. As a practical tip for the near future when using social media, spending time in immersive environments of social VR or at other social virtual crossroads, it may be helpful to recall the words of Pessoa: *"there are metaphors more real than the people who walk in the street"* [198].

On the whole, a societal debate on the difference between Type I and Type II entities may be helpful to successfully navigate the deepfake era where neither truth nor falsification nor the ability to create better explanations are lost (see Chapter 4), but where unprepared humans risk to loose grip on the world and be outmaneuvred by unscrupulous adversaries – if not equipped with a robust epistemology. In addition, from my perspective, the present-time elevated focus on intelligence – a dimension at which Type I AI is perceived to outcompete humans – instead of creativity may weaken not only the perceived agency of humanity as a whole (see also Chapter 9) but also security itself since obfuscating both the infinite creative potential of cyborgneticity that all Type II entities share and the fact that the price of security is eternal creativity [10]. If Type I AI is not specifically harnessed to augment the *creativity* of our *cyborgnet as a whole* at each network level – which is inherently always of Type II, humans risk indeed to loose control and indirectly cause catastrophic outcomes of existential dimension. However, such a loss of control would have *no link whatsoever* with the Type I AI actually becoming *qualitatively* superior to human beings. Such a Type I AI would not have the capacity to even understand what the word "destruction" (being derived from EI contexts) signifies.

To recapitulate, it is *not* intelligence that characterizes an entity as Type II but the ability to *create and understand* new EI. Even a fictive future Type *II* quantum AI (e.g. an

artificial person with quantum algorithms specifically crafted to govern certain requisite low-level computations), processing EI at unprecedented *speeds* would only represent a *quantitative* difference to present-day humanity. For non-conscious Type I AI, instead of optimizing on performance measures that relate to parameters such as "intelligence", it may be more wise to ask how the implemented AI contributes in augmenting the *creativity* of the cyborgnet(s) that utilize it. Not surprisingly, creativity cannot be quantified in the way classical key performance indicators do. Since the future of EB creation is unpredictable, the space from which one is sampling will always be incomplete leading notions of probability ad absurdum. However, that is simply part of the inescapable epistemic dizziness of Type-II-ness. As long as one is aware of the permanently required changes and updates, one can attempt to craft dynamically updatable temporary heuristical models of creativity with quantitative parameters extended by qualitative elements. This curiously leads back to moral programming [267] and augmented utilitarianism [10, 16]. Indeed, one very effective moral programming type may be cyborgnetic creativity augmentation[1] [16]. The motivation for this conjecture is threefold: 1) morality is harm-based [232], 2) to be free from harm is being secure and 3) the price of security is eternal creativity. Thereby, instead of a human-centered approach, a cyborgnetic lens may be more appropriate as explained in-depth in Chapter 3. On a final note, as can be extracted from Chapter 3 and as shortly mentioned in Chapter 10, I postulate that *we* are more than just our biological Type II substrate (i.e. *the body* including the brain) and inseparably intertwined [126] with that, *the mind*. We are also e.g. composed of all unknown and known Type I elements[2] that affect us or that we affect or imagine to affect in the factual or counterfactual past and future, also potentially including all conceivable words – i.e. obviously including "infinity". In short, we are also a *generic* cyborgnetic *template*, an infinite potential that is reliably possible in this universe. Without further commenting on it from a programming-related angle, I add that oddly or precisely not oddly, it is possible that religiously or/and spiritually inclined people would call this generic template simply *the soul*.

[1] On that view, it is noteworthy to mention that *art* can represent a form of moral programming too.

[2] Note that the thought of another Type II entity is counted as Type I since it is an *idea* and not the substrate/mind of the person directly. Generally, an entire cyborgnet is always of Type II since per definition it includes at least one physically instantiated Type II entity. For more details, see Chapter 3.

Chapter 12

Future Research

"The world is made of qubits."

David Deutsch

In any case, this is my last cyborgnetic book *primarily* aimed at EB *decryption*. For the near future, my scientific work as a cyborgnetician could focus on hidden epistemic distortions in extended reality and the metaverse but also on one practicable, low-cost and sufficiently "quantum-safe" defense strategy against hypothetical future cyberattackers equipped with early scalable Type I quantum computers willing to use those – in conjunction with suitable Type I AIs – for malicious purposes including epistemic distortions in science. (A compressed version of my first novel EB on this idea could *or* could *not* be encrypted in Appendix A.) However, the reason to target the practicability of this research idea would be purely theoretical and be linked to conjectures from cyborgnetic philosophy pertaining to new ontological distinctions in the space of (obviously physically instantiated) EI. In short, this research direction, though based on practically relevant thought experiments, would be independent of the actual practical deployment of any scalable quantum computer. In parallel, I could continue to generate visual cyborgnetic art and cyborgnetic poetry, potentially encoding equivocal EB-like seeds of counterfactual sensory-motor and cognitive-affective simulations that could but not necessarily will stimulate the creativity of the cyborgnet. Finally, I could also fill more pages of my book on cyborgnetic metaphysics having the Dutch title "So(m)nogrammaticaal - Het Lied van Tali" but written in *cyborgnettish* which is, as hinted in Chapter 8, a novel generic meta-language solely developed for purposes of EB *encryption*. However, as a cyborgnetic funambulist, one is neither committed to any EB nor to funambulism itself. As a cyborgnetic somnambulist, I might as well do none of the above and go back to sleep...

Acknowledgements

This book is dedicated to the late Ramani Jayanthi Aliman, the late Pearl Kuruneru and the late Vasantha Kuruneru. May the mini Ramani-Pearl grants further stimulate the creativity of the cyborgnet. Beyond that, I thank Leon Kester, my husband, who was the co-author of Chapter 4, was responsible for organizational matters and reviewed most other chapters. The cyborgnet thanks the second, third and fourth cyborgneticians for their future-oriented comments on the topic of *deepfake transformation in cyborgnets* which they delivered under the pseudonyms Zadhi Rangappa, Lijn Wallenstein and Cécile Leger-Zhang respectively. Being a meta-discipline, cyborgnetics can harness knowledge from an open-ended number of research areas – including *cybernetics* [22]. This book pointed out multiple ways in which humans and Type I AI could already now become *ethical regulators* of each other as envisioned by Mick Ashby [21] – whom I thank for many thought-provoking remarks a few years ago. I also thank Roman Yampolskiy for the domino-effect he unknowingly initiated in 2016 by providing helpful inputs on my early hypotheses about *cyborgization* leading from a corresponding early paper[1] [9] to this very first cyborgnetic book. I thank Judith Masthoff for facilitating access to academic literature via the account I kept at Utrecht University as a visiting scholar. Finally, I am thankful to the late Danny Frederick for providing missing pieces and novel explanations which deepened many aspects of critical rationalism and which remarkably included a solution to the *pragmatic* problem of induction [100] which Popper still faced. While unbound(ed) epistemic funambulism subsumes a diverging approach to existential questions by virtue of being a tripartite epistemic bedrock that comprises cyborgnetics, cyborgnetic philosophy and cyborgnetic art, it is the case that cyborgnetics itself profited significantly from Frederick's renewed account of critical rationalism [97].

[1] In the meantime, I improved many EBs of relevance such that I would have to thoroughly revise the mentioned paper would I ever reread it. However, I opt to go back to sleep.

Appendices

Appendix A

Potentially Encrypted EB

(See next page)

Index	Paragraph
a	AAAAAAAAAAAABCDDDDDDEEEEEEEEEEFGGHHHHHHHIIIILLMMNNNNNNNNNNNNNOOOOPRSSSSSSTTTTUUUUUUWW
b	AAAACCCDDEEEEEEGHIIIIIIILLLLMNNNOOOOOPRRRRRSSSSSSTTTUUUVWYYZ
c	AAAAABBCCDDDDEEEEEEEEEEEEFFFGHHHHHIIIIIIIKLLLLMMMMNNNNNNOOOOOOOOOOOOOPPQRRRRRRRSSSSSSTTTTTTTTTTTUUUWWWY
d	AAAAAAABCCCCEEEEGHIIIIIKLLLMMNNNNNNNOOOOOOOPPRRRSSSTTTTTUUXYY
e	AAAAAAAAAAAAAAAAAAABBBBBBCCCCCCCCCCCDDDDEEEEEEEEEEEEEEFFFFFFFGGGGHHHHIIIIIIIIIIKKKLLLLLMMMMMMMNNNNNNNNOOOOOOOOOPPPRRRRRRRRRSSSSSSSSTTTTTTTTTTTTTUWXYYY
f	AAAAAAAAAAAAAAAAAAAAAAAABBBBBBBCCCCCCCCCCCDDDDDDEEEEEEEEEEEEEEFFFFFFGGGGGGHHHHHIIIIIIIIIIIIIIIKKKKKLLLLLLLLLMMMMMMMMMMMMMNNNNNNNNNNNNNOOOOOOOOOOOOOOOPRRRRRRRRRRRRSSSSSSSSSSTTTTTTTTTTTTTTTTTUUVVWWWXXY
g	AAAABCCDDDDEEEEEEEEEEEEEEEEFFGHHHHHHIILMMMNNNNNOOOORRRRRRSSSSSSTTTTTTTUUUVVWYY
h	AAAAACCDDDEEEEEEEEEGHHIIIIIILLLNNNNOOOOPRRRRRRSSSTTTTTTUVWWWZ
i	AAAAAAAAAAAAAAAAAAAAABBBCCCCCCCCCCCCCDDDEEEEEEEEEEEEEEEEEEEFFFFGGGGGHHHHHHIIIIIIIIIIIIIIIIIIKLLLLLLMMMMMMMMMNNNNNNNNNNNNNNNOOOOOOOOOOOOOOPPRRRRRRRRRSSSSSSSSSSSSSTTTTTTTTTTTTTTTTTTUUUUVWWXYY
j	AAAAACCCDDEEEEEEEGHIIILMNNNOOOPRRRSSSSSTTTUVW
k	AAAAAAAABDEEEEEEEEEEEEFFHHHIIIIIIIIIIIIIIKKLLLLLMNNNNNNNNNNOOOOOOOOOOPPPRRRRRSSSSSSSTTTTTTTTTTTTTUUUUVVWWWYYYZ
l	AAAAAAAAACDDEEEEEEEEEEEEEFFGGGHHHHHHHIIIIIIIIIILLLLMMNNNNNNNNNOOOPPRRRRRRSSSSSTTTTTTTTTTTUX
m	AAAAAAAAAAAAAABCCCCCCCCCDDDDDDDDDDEEEEEEEEEEEEEEEEEEEEFFFFFGGGGGGGGHHHHHHHHHIIIIIIIIIIIIIIIIJMMMMNNNNNNNNNNNNNNNNNNNNNOOOOOOOOPPPQRRRRRRRRSSSSSSSTTTTTTTTTTTTTTTUUUUVWWWY
n	AAAAAAAABBBBBCCCCCCCCCDDDDDEEEEEEEEEEEEEEEEEEEEEEEEFFFGHHHHHHHIIIIIIIIIIIIIIIKLLLLMNNNNNNNNNNNNNNNOOOOOOOOOOOPPPPPPPRRRRRRSSSSSSTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTUUUUVWWYYYYYYYYY
o	AAAAAAAAAAAAAAAAAAAAAAAABBBCCCCCCCCCCDDEEEEEEEEEEEEEEFFGGGHHHHHIIIIIIIIIIIIIIIIKKKKLLLLLLLLLMMMMMMNNNNNNNNNNNNNNNOOOOOOOOOOOOOOOPPPPPPRRRRRRRRRRSSSSSSSTTTTTTTTTTTTTTTTUVWWXXXXYYYY

Encrypted chain?	Yes	No

5-letter-guess (if yes)					

If yes: **Why** these 5 letters (and not 5 of the remaining 10 others)?	If no: **Why** not?

Bibliography

[1] A. Abbott. Brain study probes primate 'software'. *Nature*, 565(7740):410–411, 2019.

[2] E.-S. Abd-Elaal, S. H. Gamage, J. E. Mills, et al. Artificial intelligence is a tool for cheating academic integrity. In *30th Annual Conference for the Australasian Association for Engineering Education (AAEE 2019): Educators Becoming Agents of Change: Innovate, Integrate, Motivate*, page 397. Engineers Australia, 2019.

[3] S. Abdelnabi and M. Fritz. Adversarial watermarking transformer: Towards tracing text provenance with data hiding. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 121–140. IEEE, 2021.

[4] A. Abdibayev, D. Chen, H. Chen, D. Poluru, and V. Subrahmanian. Using Word Embeddings to Deter Intellectual Property Theft through Automated Generation of Fake Documents. *ACM Transactions on Management Information Systems (TMIS)*, 12(2):1–22, 2021.

[5] Agence France-Presse. FBI has 1,000 investigations into Chinese intellectual property theft, director Christopher Wray says, calling China the most severe counter-intelligence threat to US. `https://www.scmp.com/news/china/article/3019829/fbi-has-1000-probes-chinese-intellectual-property-theft-director`, 2019. Online; accessed 26-June-2021.

[6] I. Aggarwal, A. W. Woolley, C. F. Chabris, and T. W. Malone. The impact of cognitive style diversity on implicit learning in teams. *Frontiers in psychology*, 10:112, 2019.

[7] L. F. Agnati, P. Barlow, R. Ghidoni, D. O. Borroto-Escuela, D. Guidolin, and K. Fuxe. Possible genetic and epigenetic links between human inner speech, schizophrenia and altruism. *Brain research*, 1476:38–57, 2012.

[8] B. Alderson-Day and C. Fernyhough. Inner speech: development, cognitive functions, phenomenology, and neurobiology. *Psychological bulletin*, 141(5):931, 2015.

[9] N.-M. Aliman. Malevolent cyborgization. In *International Conference on Artificial General Intelligence*, pages 188–197. Springer, 2017.

[10] N.-M. Aliman. *Hybrid Cognitive-Affective Strategies for AI Safety*. PhD thesis, Utrecht University, 2020.

[11] N.-M. Aliman, P. Elands, W. Hürst, L. Kester, K. R. Thórisson, P. Werkhoven, R. Yampolskiy, and S. Ziesche. Error-correction for AI safety. In *International Conference on Artificial General Intelligence*, pages 12–22. Springer, 2020.

[12] N.-M. Aliman and L. Kester. Extending socio-technological reality for ethics in artificial intelligent systems. In *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 275–2757. IEEE, 2019.

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

[14] N.-M. Aliman and L. Kester. Facing Immersive "Post-Truth" in AIVR? *Philosophies*, 5(4):45, 2020.

[15] 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.

[16] N.-M. Aliman and L. Kester. Moral programming – Crafting a flexible heuristic moral meta-model for meaningful AI control in pluralistic societies. In *Moral Design and Technology*, page to appear. Wernaart, Bart, 2022.

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

[18] B. Ambrosino. How and why did religion evolve. `https://www.bbc.com/future/article/20190418-how-and-why-did-religion-evolve`, 2019. BBC Future; accessed 30-May-2021.

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

[20] J. R. Andrews-Hanna. The brain's default network and its adaptive role in internal mentation. *The Neuroscientist*, 18(3):251–270, 2012.

[21] M. Ashby. Ethical regulators and super-ethical systems. *Systems*, 8(4):53, 2020.

[22] W. R. Ashby. *An introduction to cybernetics*. Chapman & Hall Ltd, 1961.

[23] A. Ashkenazy and S. Zini. Attacking Machine Learning – The Cylance Case Study . `https://skylightcyber.com/2019/07/18/cylance-i-kill-you/Cylance%20-`

%20Adversarial%20Machine%20Learning%20Case%20Study.pdf, 2019. Skylight; accessed 24-May-2020.

[24] S. Atzil, W. Gao, I. Fradkin, and L. F. Barrett. Growing a social brain. *Nature Human Behaviour*, 2(9):624–636, 2018.

[25] E. Bandini, A. Motes-Rodrigo, W. Archer, T. Minchin, H. Axelsen, R. A. Hernandez-Aguilar, S. P. McPherron, and C. Tennie. Naïve, unenculturated chimpanzees fail to make and use flaked stone tools. *Open Research Europe*, 1:20, 2021.

[26] L. Barham and D. Everett. Semiotics and the Origin of Language in the Lower Palaeolithic. *Journal of Archaeological Method and Theory*, pages 1–45, 2020.

[27] I. Baris and Z. Boukhers. ECOL: Early Detection of COVID Lies Using Content, Prior Knowledge and Source Information. *arXiv preprint arXiv:2101.05499*, 2021.

[28] L. F. Barrett. The conceptual act theory: A précis. *Emotion review*, 6(4):292–297, 2014.

[29] L. F. Barrett. Functionalism cannot save the classical view of emotion. *Social Cognitive and Affective Neuroscience*, 12(1):34–36, 2017.

[30] L. F. Barrett. *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt, 2017.

[31] L. F. Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1):1–23, 2017.

[32] 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.

[33] M. Benz and D. Chatterjee. Calculated risk? A cybersecurity evaluation tool for SMEs. *Business Horizons*, 63(4):531–540, 2020.

[34] M. J. Beran and L. A. Heimbauer. A longitudinal assessment of vocabulary retention in symbol-competent chimpanzees (Pan troglodytes). *PloS one*, 10(2):e0118408, 2015.

[35] S. L. Bernal, A. H. Celdrán, G. M. Pérez, M. T. Barros, and S. Balasubramaniam. Security in Brain-Computer Interfaces: State-of-the-Art, Opportunities, and Future Challenges. *ACM Computing Surveys (CSUR)*, 54(1):1–35, 2021.

[36] M. Bettoni. The Yerkish language. From operational methodology to chimpanzee communication. Constructivist Foundations 2 (2–3): 32–38, 2007.

[37] L. F. Bicchierai. Hacked Toy Company VTech: Let Us Monitor Your Houses. `https://www.vice.com/en/article/xygxxw/hacked-toy-company-vtech-let-us-monitor-your-house`, 2016. VICE; accessed 08-September-2021.

[38] Y. Bigman, K. Gray, A. Waytz, M. Arnestad, and D. Wilson. Algorithmic discrimination causes less moral outrage than human discrimination. *PsyArXiv*, 2020.

[39] L. Bilge and T. Dumitraş. Before we knew it: an empirical study of zero-day attacks in the real world. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 833–844, 2012.

[40] D. Bolis, J. Balsters, N. Wenderoth, C. Becchio, and L. Schilbach. Beyond autism: Introducing the dialectical misattunement hypothesis and a Bayesian account of intersubjectivity. *Psychopathology*, 50(6):355–372, 2017.

[41] T. Bonaci, R. Calo, and H. J. Chizeck. App stores for the brain: Privacy & security in Brain-Computer Interfaces. In *2014 IEEE International Symposium on Ethics in Science, Technology and Engineering*, pages 1–7. IEEE, 2014.

[42] D. Boneh, A. J. Grotto, P. McDaniel, and N. Papernot. How relevant is the Turing test in the age of sophisbots? *IEEE Security & Privacy*, 17(6):64–71, 2019.

[43] N. Bostrom. Strategic implications of openness in AI development. *Global policy*, 8(2):135–148, 2017.

[44] N. Boucher, I. Shumailov, R. Anderson, and N. Papernot. Bad Characters: Imperceptible NLP Attacks. *arXiv preprint arXiv:2106.09898*, 2021.

[45] K. E. Brakke and E. S. Savage-Rumbaugh. The development of language skills in bonobo and chimpanzee: I. Comprehension. *Language & Communication*, 1995.

[46] H. M. Bronte-Stewart, M. N. Petrucci, J. J. O'Day, M. F. Afzal, J. E. Parker, Y. M. Kehnemouyi, K. B. Wilkins, G. C. Orthlieb, and S. L. Hoffman. Perspective: Evolution of Control Variables and Policies for Closed-Loop Deep Brain Stimulation for Parkinson's Disease Using Bidirectional Deep-Brain-Computer Interfaces. *Frontiers in Human Neuroscience*, 14:353, 2020.

[47] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[48] 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.

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

[50] T.-C. Bui, V.-D. Le, H.-T. To, and S. K. Cha. Generative Pre-training for Paraphrase Generation by Representing and Predicting Spans in Exemplars. In *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 83–90. IEEE, 2021.

[51] J. Burden and J. Hernández-Orallo. Exploring AI Safety in Degrees: Generality, Capability and Control. In *SafeAI@ AAAI*, pages 36–40, 2020.

[52] 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.

[53] Y. Cai, W. van Joolingen, and K. Veermans. Virtual and Augmented Reality, Simulation and Serious Games for Education, 2021.

[54] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

[55] P. Casey, I. Baggili, and A. Yarramreddy. Immersive virtual reality attacks and the human joystick. *IEEE Transactions on Dependable and Secure Computing*, 2019.

[56] T. Chakraborty, S. Jajodia, J. Katz, A. Picariello, G. Sperli, and V. Subrahmanian. FORGE: A fake online repository generation engine for cyber deception. *IEEE Transactions on Dependable and Secure Computing*, 2019.

[57] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. Ponde, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[58] W.-K. Chen. *Applied Graph Theory: Graphs and Electrical Networks*. Elsevier, 2014.

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

[60] S. Chikhale and V. Gohad. Multidimensional Construct About the Robot Citizenship Law's in Saudi Arabia. *International Journal of Innovative Research and Advanced Studies (IJIRAS)*, 5(1):106–108, 2018.

[61] N. Chomsky. Aspects of the theory of syntax (Cambridge, Mass.). *Multilingual Matters: MIT Press*, 1965.

[62] N. Chomsky. *Aspects of the Theory of Syntax*, volume 11. MIT press, 2014.

[63] M. H. Christiansen and N. Chater. The language faculty that wasn't: a usage-based account of natural language recursion. *Frontiers in Psychology*, 6:1182, 2015.

[64] F. Ciardo, F. Beyer, D. De Tommaso, and A. Wykowska. Attribution of intentional agency towards robots reduces one's own sense of agency. *Cognition*, 194:104109, 2020.

[65] A. Ciaunica, A. Constant, H. Preissl, and A. Fotopoulou. The First Prior: from Co-Embodiment to Co-Homeostasis in Early Life, Jan 2021.

[66] R. V. Clarke. Technology, criminology and crime science. *European Journal on Criminal Policy and Research*, 10(1):55–63, 2004.

[67] Z. A. Collier and J. Sarkis. The zero trust supply chain: Managing supply chain risk in the absence of trust. *International Journal of Production Research*, pages 1–16, 2021.

[68] J. d. A. da Cruz and S. Pedron. Cyber Mercenaries: A New Threat to National Security. *International Social Science Review*, 96(2):3, 2020.

[69] K. G. Davis. Echoes of Language Development: 7 Facts About Echolalia for SLPs. *Leader Live*, 2017.

[70] J. A. De Guzman, K. Thilakarathna, and A. Seneviratne. Security and privacy approaches in mixed reality: A literature survey. *ACM Computing Surveys (CSUR)*, 52(6):1–37, 2019.

[71] N. Dehouche. Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3). *Ethics in Science and Environmental Politics*, 21:17–23, 2021.

[72] W. S. DeKeseredy. *Contemporary critical criminology*. Routledge, 2010.

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

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

[75] 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.

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

[77] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[78] A. Dietrich. Where in the brain is creativity: a brief account of a wild-goose chase. *Current Opinion in Behavioral Sciences*, 27:36–39, 2019.

[79] S. Dolcos and D. Albarracin. The inner speech of behavioral regulation: Intentions and task performance strengthen when you talk to yourself as a You. *European Journal of Social Psychology*, 44(6):636–642, 2014.

[80] L. Dubreuil and S. Savage-Rumbaugh. Dialogues on the Human Ape, 2018.

[81] H. Else et al. 'Tortured phrases' give away fabricated research papers. *Nature*, 596(7872):328–329, 2021.

[82] M. I. Eren, S. J. Lycett, and M. Tomonaga. Underestimating Kanzi? Exploring Kanzi-Oldowan comparisons in light of recent human stone tool replication. *Evolutionary Anthropology: Issues, News, and Reviews*, 29(6):310–316, 2020.

[83] D. Everett. *How language began: The story of humanity's greatest invention*. Profile Books, 2017.

[84] D. L. Everett. What does Pirahã grammar have to teach us about human language and the mind? *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(6):555–563, 2012.

[85] D. L. Everett. *Dark matter of the mind: the culturally articulated unconscious*. University of Chicago Press, 2016.

[86] D. L. Everett. Grammar came later: triality of patterning and the gradual evolution of language. *Journal of Neurolinguistics*, 43:133–165, 2017.

[87] D. L. Everett. The role of culture in language and cognition. *Language and Linguistics Compass*, 12(11):e12304, 2018.

[88] D. Fallis. The Epistemic Threat of Deepfakes. *Philosophy & Technology*, pages 1–21, 2020.

[89] E. Ferrara and Z. Yang. Measuring emotional contagion in social media. *PloS one*, 10(11):e0142390, 2015.

[90] A. Fickinger, S. Zhuang, D. Hadfield-Menell, and S. Russell. Multi-principal assistance games. *arXiv preprint arXiv:2007.09540*, 2020.

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

[92] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, et al. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4):689–707, 2018.

[93] R. A. Foley. Mosaic evolution and the pattern of transitions in the hominin lineage. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1698):20150244, 2016.

[94] T. Foltỳnek, T. Ruas, P. Scharpf, N. Meuschke, M. Schubotz, W. Grosky, and B. Gipp. Detecting Machine-obfuscated Plagiarism. In *International Conference on Information*, pages 816–827. Springer, 2020.

[95] P. Fransson, B. Skiöld, S. Horsch, A. Nordell, M. Blennow, H. Lagercrantz, and U. Åden. Resting-state networks in the infant brain. *Proceedings of the National Academy of Sciences*, 104(39):15531–15536, 2007.

[96] D. Frederick. Are Institutions Created by Collective Acceptance? *The Journal of Value Inquiry*, pages 1–13, 2019.

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

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

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

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

[101] E. Fridland. Do as I say and as I do: imitation, pedagogy, and cumulative culture. *Mind & Language*, 33(4):355–377, 2018.

[102] W. Gao, J. H. Gilmore, D. Shen, J. K. Smith, H. Zhu, and W. Lin. The synchronization within and interaction between the default and dorsal attention networks in early infancy. *Cerebral cortex*, 23(3):594–603, 2013.

[103] T. G. García-Micó. Electronic Personhood: A Tertium Genus for Smart Autonomous Surgical Robots? In *Algorithmic Governance and Governance of Algorithms*, pages 87–108. Springer, 2021.

[104] M. Gault. Six Reasons why Encryption isn't working . `https://guardtime.com/blog/6-reasons-why-encryption-isnt-working`, 2021. Guardtime; accessed 12-August-2021.

[105] M. Ghallab. Responsible AI: requirements and challenges. *AI Perspectives*, 1(1):1–7, 2019.

[106] G. Goehring. Keplers Solutions to Galileos Anagrams. *Journal of the British Astronomical Association*, 92:41, 1981.

[107] A. Goranson, P. Sheeran, J. Katz, and K. Gray. Doctors are seen as Godlike: Moral typecasting in medicine. *Social Science & Medicine*, 258:113008, 2020.

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

[109] GPT-3. A robot wrote this entire article. Are you scared yet, human? *The Guardian*, 2020.

[110] C. Grassley. Grassley on Chinese Espionage: It's called cheating. And it's only getting worse. `https://www.judiciary.senate.gov/grassley-on-chinese-espionage-its-called-cheating_and-its-only-getting-worse`, 2019. CNBC; accessed 26-June-2021.

[111] V. Grech. Fake news and post-truth pronouncements in general and in early human development. *Early Human Development*, 115:118–120, 2017.

[112] D. L. Greenberg and M. Verfaellie. Interdependence of episodic and semantic memory: evidence from neuropsychology. *Journal of the International Neuropsychological Society: JINS*, 16(5):748, 2010.

[113] J. Greenberg, S. Solomon, and J. Arndt. A Basic but Uniquely Human Motivation. *Handbook of Motivation Science*, page 114, 2013.

[114] 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.

[115] 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.

[116] D. Heaney. This OpenAI GPT-3 Powered Demo Is A Glimpse Of NPCs In The Future. `https://uploadvr.com/modbox-gpt3-ai-npc-demo/`, 2021. UploadVR; accessed 08-September-2021.

[117] M. G. Henriksen, J. Parnas, and D. Zahavi. Thought insertion and disturbed for-meness (minimal selfhood) in schizophrenia. *Consciousness and cognition*, 74:102770, 2019.

[118] S. Herculano-Houzel. *The human advantage: a new understanding of how our brain became remarkable*. MIT Press, 2016.

[119] J. Hernández-Orallo, F. Martínez-Plumed, S. Avin, J. Whittlestone, and S. Ó. hÉigeartaigh. AI Paradigms and AI Safety: Mapping Artefacts and Techniques to Safety Issues. In *ECAI*, 2020.

[120] E. Herrmann, J. Call, M. V. Hernández-Lloreda, B. Hare, and M. Tomasello. Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *science*, 317(5843):1360–1366, 2007.

[121] P. Heyvaert, B. De Meester, A. Dimou, and R. Verborgh. Rule-driven inconsistency resolution for knowledge graph generation rules. *Semantic Web*, 10(6):1071–1086, 2019.

[122] A. Hill and S. Ward. Origin of the Hominidae: the record of African large hominoid evolution between 14 My and 4 My. *American Journal of Physical Anthropology*, 31(S9):49–83, 1988.

[123] P. Hillyard and S. Tombs. Social harm and zemiology. *The Oxford handbook of criminology*, pages 284–305, 2017.

[124] S. S. Ho, T. J. Goh, and Y. W. Leung. Let's nab fake science news: Predicting scientists' support for interventions using the influence of presumed media influence model. *Journalism*, page 1464884920937488, 2020.

[125] E. Hoel. The overfitted brain: Dreams evolved to assist generalization. *Patterns*, 2(5):100244, 2021.

[126] K. Hoemann and L. Feldman Barrett. Concepts dissolve artificial boundaries in the study of emotion and cognition, uniting body, brain, and mind. *Cognition and Emotion*, 33(1):67–76, 2019.

[127] H. Hopf, A. Krief, G. Mehta, and S. A. Matlin. Fake science and the knowledge crisis: ignorance can be fatal. *Royal Society open science*, 6(5):190161, 2019.

[128] S. Houde, V. Liao, J. Martino, M. Muller, D. Piorkowski, J. Richards, J. Weisz, and Y. Zhang. Business (mis) Use Cases of Generative AI. *arXiv preprint arXiv:2003.07679*, 2020.

[129] Q. Hu, Y. Lu, Z. Pan, Y. Gong, and Z. Yang. Can AI artifacts influence human cognition? The effects of artificial autonomy in intelligent personal assistants. *International Journal of Information Management*, 56:102250, 2021.

[130] S. Izawa, S. Chowdhury, T. Miyazaki, Y. Mukai, D. Ono, R. Inoue, Y. Ohmura, H. Mizoguchi, K. Kimura, M. Yoshioka, et al. REM sleep–active MCH neurons are involved in forgetting hippocampus-dependent memories. *Science*, 365(6459):1308–1313, 2019.

[131] O. L. Jacobs, K. Gazzaz, and A. Kingstone. Mind the robot! Variation in attributions of mind to a wide set of real and fictional robots. *International Journal of Social Robotics*, pages 1–9, 2021.

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

[133] K. Jensen, J. Call, and M. Tomasello. Chimpanzees are rational maximizers in an ultimatum game. *science*, 318(5847):107–109, 2007.

[134] A. Jobin, M. Ienca, and E. Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019.

[135] C. Joshi. Transformers are Graph Neural Networks. `https://www.nadishamarie.jimdo.com/clipboard/`, 2020. The Gradient; accessed 26-August-2021.

[136] G. P. T. Jr, E. X. Note, M. S. Spellchecker, and R. Yampolskiy. When Should Co-Authorship Be Given to AI? `https://philarchive.org/archive/GPTWSCv1`, 2020. Unpublished, PhilArchive; accessed 08-November-2020.

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

[138] I. Kalpokas. Problematising reality: the promises and perils of synthetic media. *SN Social Sciences*, 1(1):1–11, 2021.

[139] S. Kang, C. Molinaro, A. Pugliese, and V. Subrahmanian. Randomized Generation of Adversary-aware Fake Knowledge Graphs to Combat Intellectual Property Theft. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4155–4163, 2021.

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

[141] M. Kianpour. Socio-Technical Root Cause Analysis of Cyber-enabled Theft of the US Intellectual Property–The Case of APT41. *arXiv preprint arXiv:2103.04901*, 2021.

[142] H. Kiilavuori, V. Sariola, M. J. Peltola, and J. K. Hietanen. Making eye contact with a robot: Psychophysiological responses to eye contact with a human and with a humanoid robot. *Biological Psychology*, 158:107989, 2021.

[143] T. Kim and H. Song. How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics*, 61:101595, 2021.

[144] J. Kindervag. Build security into your network's DNA: The zero trust network architecture. *Forrester Research Inc*, pages 1–26, 2010.

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

[146] E. Koechlin. Frontal pole function: what is specifically human? *Trends in cognitive sciences*, 15(6):241, 2011.

[147] E. Kolthoff and J. Janssen. *Basisboek criminologie*. Boom Lemma, 2020.

[148] M. W. Kranenbarg, T. J. Holt, and J. van der Ham. Don't shoot the messenger! A criminological and computer science perspective on coordinated vulnerability disclosure. *Crime Science*, 7(1):1–9, 2018.

[149] G. Laban, V. Morrison, and E. S. Cross. Let's Talk About It! Subjective and Objective Disclosures to Social Robots. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 328–330, 2020.

[150] T. Lackorzynski and S. Koepsell. " Hello Barbie"-Hacker Toys in a World of Linked Devices. In *Broadband Coverage in Germany; 11. ITG-Symposium*, pages 1–7. VDE, 2017.

[151] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.

[152] J. LeDoux. *The deep history of ourselves: The four-billion-year story of how we got conscious brains*. Penguin Books, 2020.

[153] A. K. Lee and M. A. Wilson. Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron*, 36(6):1183–1194, 2002.

[154] S. Lee, N. Lee, and Y. J. Sah. Perceiving a mind in a Chatbot: effect of mind perception and social cues on co-presence, closeness, and intention to use. *International Journal of Human–Computer Interaction*, 36(10):930–940, 2020.

[155] D. Lefkeli, Y. Ozbay, Z. Gürhan-Canli, and T. Eskenazi. Competing with or Against Cozmo, the Robot: Influence of Interaction Context and Outcome on Mind Perception. *International Journal of Social Robotics*, pages 1–10, 2020.

[156] J. Leike, M. Martic, V. Krakovna, P. A. Ortega, T. Everitt, A. Lefrancq, L. Orseau, and S. Legg. AI safety gridworlds. *arXiv preprint arXiv:1711.09883*, 2017.

[157] J. M. Lemnitzer. Why cybersecurity insurance should be regulated and compulsory. *Journal of Cyber Policy*, pages 1–19, 2021.

[158] P. M. Leonardi. When flexible routines meet flexible technologies: Affordance, constraint, and the imbrication of human and material agencies. *MIS quarterly*, pages 147–167, 2011.

[159] H. M. Lewis and K. N. Laland. Transmission fidelity is the key to the build-up of cumulative culture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1599):2171–2180, 2012.

[160] P. A. Lewis, G. Knoblich, and G. Poe. How memory replay in sleep boosts creative problem-solving. *Trends in cognitive sciences*, 22(6):491–503, 2018.

[161] G. Lima, M. Cha, C. Jeon, and K. Park. Explaining the punishment gap of AI and robots. *arXiv e-prints*, pages arXiv–2003, 2020.

[162] H. Ling and E. Björling. Sharing stress with a robot: what would a robot say? *Human-Machine Communication*, 1, 2020.

[163] W. R. Liu, Q. Cao, and Y. Cai. Serious Game Design for Virtual Dolphin-Assisted Learning. *When VR Serious Games Meet Special Needs Education: Research, Development and Their Applications*, pages 97–112, 2021.

[164] H. Lyn. Mental representation of symbols as revealed by vocabulary errors in two bonobos (Pan paniscus). *Animal Cognition*, 10(4):461–475, 2007.

[165] H. Lyn, B. Franks, and E. S. Savage-Rumbaugh. Precursors of morality in the use of the symbols "good" and "bad" in two bonobos (Pan paniscus) and a chimpanzee (Pan troglodytes). *Language & Communication*, 28(3):213–224, 2008.

[166] H. Lyn, P. Greenfield, and S. Savage-Rumbaugh. The development of representational play in chimpanzees and bonobos: Evolutionary implications, pretense, and the role of interspecies communication. *Cognitive Development*, 21(3):199–213, 2006.

[167] H. Lyn, P. M. Greenfield, and E. S. Savage-Rumbaugh. Semiotic combinations in Pan: A comparison of communication in a chimpanzee and two bonobos. *First Language*, 31(3):300–325, 2011.

[168] H. Lyn, P. M. Greenfield, S. Savage-Rumbaugh, K. Gillespie-Lynch, and W. D. Hopkins. Nonhuman primates do declare! A comparison of declarative symbol

and gesture use in two children, two bonobos, and a chimpanzee. *Language & communication*, 31(1):63–74, 2011.

[169] H. Lyn and E. S. Savage-Rumbaugh. Observational word learning in two bonobos (Pan paniscus): Ostensive and non-ostensive contexts. *Language & Communication*, 20(3):255–273, 2000.

[170] H. Lyn and S. Savage-Rumbaugh. The use of emotion symbols in language-using apes. In *Emotions of Animals and Humans*, pages 113–127. Springer, 2012.

[171] T. Mahlangu, S. January, T. Mashiane, M. Dlamini, S. Ngobeni, N. Ruxwana, and S. Tzu. 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.

[172] A. Makri. Give the public the tools to trust scientists. *Nature News*, 541(7637):261, 2017.

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

[174] B. F. Malle and D. Ullman. A multidimensional conception and measure of human-robot trust. In *Trust in Human-Robot Interaction*, pages 3–25. Elsevier, 2021.

[175] F. Manzi, G. Peretti, C. Di Dio, A. Cangelosi, S. Itakura, T. Kanda, H. Ishiguro, D. Massaro, and A. Marchetti. A robot is not worth another: exploring children's mental state attribution to different humanoid robots. *Frontiers in Psychology*, 11:2011, 2020.

[176] D. U. Martin, M. I. MacIntyre, C. Perry, G. Clift, S. Pedell, and J. Kaufman. Young children's indiscriminate helping behavior toward a humanoid robot. *Frontiers in psychology*, 11:239, 2020.

[177] W. MathWorld. NP-Hard Problem . `https://mathworld.wolfram.com/NP-HardProblem.html`, 2021. Online; accessed 26-June-2021.

[178] T. Matsuzawa. Symbolic representation of number in chimpanzees. *Current opinion in neurobiology*, 19(1):92–98, 2009.

[179] S. McGregor. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *arXiv preprint arXiv:2011.08512*, 2020.

[180] H. M. McHenry. Fossils and the mosaic nature of human evolution. *Science*, 190(4213):425–431, 1975.

[181] MIT Open Learning. Tackling the misinformation epidemic with "In Event of Moon Disaster" . https://news.mit.edu/2020/mit-tackles-misinformation-in-event-of-moon-disaster-0720, 2020. MIT News; accessed 11-October-2020.

[182] S. Moussawi and R. Benbunan-Fich. The effect of voice and humour on users' perceptions of personal intelligent agents. *Behaviour & Information Technology*, pages 1–24, 2020.

[183] J. Naughton. What is 'technology'. *Teaching technology*, pages 7–12, 1994.

[184] L. H. Newman. Apple's Ransomware Mess Is the Future of Online Extortion. https://www.wired.com/story/apple-ransomware-attack-quanta-computer/, 2021. Wired; accessed 26-June-2021.

[185] D. Nguyen, D. Liebowitz, S. Nepal, and S. Kanhere. HoneyCode: Automating Deceptive Software Repositories with Deep Generative Models. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, page 6945, 2021.

[186] M. Nye, M. H. Tessler, J. B. Tenenbaum, and B. M. Lake. Improving Coherence and Consistency in Neural Sequence Models with Dual-System, Neuro-Symbolic Reasoning. *arXiv preprint arXiv:2107.02794*, 2021.

[187] S. S. ÓhÉigeartaigh, J. Whittlestone, Y. Liu, Y. Zeng, and Z. Liu. Overcoming barriers to cross-cultural cooperation in AI ethics and governance. *Philosophy & Technology*, 33(4):571–593, 2020.

[188] M. Orabi, D. Mouheb, Z. Al Aghbari, and I. Kamel. Detection of Bots in Social Media: A Systematic Review. *Information Processing & Management*, 57(4):102250, 2020.

[189] B. Y. Ozkan, S. van Lingen, and M. Spruit. The Cybersecurity Focus Area Maturity (CYSFAM) Model. *Journal of Cybersecurity and Privacy*, 1(1):119–139, 2021.

[190] N. Park, A. Kan, X. L. Dong, T. Zhao, and C. Faloutsos. Estimating node importance in knowledge graphs using graph neural networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 596–606, 2019.

[191] K. Pearlman. Virtual Reality Brings Real Risks: Are We Ready? *USENIX Enigma 2020*, 2020.

[192] J. Pedersen. Apes in conversation: The role of the human interlocutor. *Language & Communication*, 50:1–11, 2016.

[193] J. Pedersen and W. M. Fields. Aspects of repetition in bonobo–human conversation: creating cohesion in a conversation between species. *Integrative Psychological and Behavioral Science*, 43(1):22–41, 2009.

[194] C. S. Peirce. *Peirce on signs: Writings on semiotic*. UNC Press Books, 1991.

[195] C. S. Peirce, L. V. Welby, V. A. M. L. Stuart-Wortley, and V. L. Welby. *Semiotic and Significs: The Correspondence Between Charles S. Peirce and Lady Victoria Welby*. Indiana University Press, 1977.

[196] S. A. Pemberton. *Harmful societies: Understanding social harm*. Policy Press, 2016.

[197] J. Perez-Osorio and A. Wykowska. Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology*, 33(3):369–395, 2020.

[198] F. Pessoa. *The book of disquiet*. Penguin UK, 2002.

[199] B. E. Pfeiffer and D. J. Foster. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, 497(7447):74–79, 2013.

[200] F. Pistono and R. V. Yampolskiy. Unethical Research: How to Create a Malevolent Artificial Intelligence. *arXiv e-prints*, pages arXiv–1605, 2016.

[201] V. Pitardi and H. R. Marriott. Alexa, she's not human but... Unveiling the drivers of consumers' trust in voice-based artificial intelligence. *Psychology & Marketing*, 38(4):626–642, 2021.

[202] K. Popper. *In search of a better world: Lectures and essays from thirty years*. Psychology Press, 1996.

[203] K. Popper. *The logic of scientific discovery*. Routledge, 2005.

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

[205] J. Prier. Commanding the trend: Social media as information warfare. *Strategic Studies Quarterly*, 11(4):50–85, 2017.

[206] R. Pryluk, Y. Kfir, H. Gelbard-Sagiv, I. Fried, and R. Paz. A tradeoff in the neural code across regions and species. *Cell*, 176(3):597–609, 2019.

[207] G. K. Pullum. Theorizing about the syntax of human language: a radical alternative to generative formalisms. *Cadernos de Linguística*, 1(1):1–33, 2020.

[208] R. Q. Quiroga. No Pattern Separation in the Human Hippocampus. *Trends in Cognitive Sciences*, 2020.

[209] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[210] I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 145–151, 2020.

[211] P. Ranade, A. Piplai, S. Mittal, A. Joshi, and T. Finin. Generating Fake Cyber Threat Intelligence Using Transformer-Based Models. *arXiv preprint arXiv:2102.04351*, 2021.

[212] T. Randhavane, A. Bera, K. Kapsaskis, R. Sheth, K. Gray, and D. Manocha. EVA: Generating emotional behavior of virtual agents using expressive features of gait and gaze. In *ACM symposium on applied perception 2019*, pages 1–10, 2019.

[213] R. Razdan. Temple Grandin, Elon Musk And The Interesting Parallels Between Autonomous Vehicles and Autism. `https://www.forbes.com/sites/rahulrazdan/2020/06/07/temple-grandin-elon-musk-and-the-interesting-parallels-between-autonomous-vehicles-and-autism/`, 2020. Forbes; accessed 04-June-2021.

[214] A. Reynolds and D. Lewis. Teams solve problems faster when they're more cognitively diverse. *Harvard Business Review*, 30, 2017.

[215] G. Richards, S. Baron-Cohen, H. Stokes, V. Warrier, B. Mellor, E. Winspear, J. Davies, L. Gee, and J. Galvin. Assortative Mating, Autistic Traits, Empathizing, and Systemizing. *bioRxiv*, 2020.

[216] R. W. Rieber and A. S. Carton. The collected works of LS Vygotsky. *Problems of general psychology*, 1:325–339, 1987.

[217] D. Robert and M. Dufresne. *Actor-network theory and crime studies: Explorations in science and technology*. Routledge, 2016.

[218] E. Rosenbaum. 1 in 5 corporations say China has stolen their IP within the last year: CNBC CFO survey. `https://www.scmp.com/news/china/article/3019829/fbi-has-1000-probes-chinese-intellectual-property-theft-director`, 2019. CNBC; accessed 26-June-2021.

[219] S. B. Rosenthal. Semiotic and Significs: The Correspondence between Charles S. Peirce and Lady Victoria Welby. *Journal of the History of Philosophy*, 17(4):487–487, 1979.

[220] J. L. Russell, H. Lyn, J. A. Schaeffer, and W. D. Hopkins. The role of socio-communicative rearing environments in the development of social and physical cognition in apes. *Developmental science*, 14(6):1459–1470, 2011.

[221] M. Sahlgren and F. Carlsson. The Singleton Fallacy: Why Current Critiques of Language Models Miss the Point. *arXiv preprint arXiv:2102.04310*, 2021.

[222] P. Sándor, S. Szakadát, K. Kertész, and R. Bódizs. Content analysis of 4 to 8 year-old children's dream reports. *Frontiers in psychology*, 6:534, 2015.

[223] R. Satter. Experts: Spy used AI-generated face to connect with targets. `https://apnews.com/article/bc2f19097a4c4fffaa00de6770b8a60d`, 2019. AP News; accessed 04-August-2020.

[224] E. S. Savage-Rumbaugh, J. Murphy, R. A. Sevcik, K. E. Brakke, S. L. Williams, D. M. Rumbaugh, and E. Bates. Language comprehension in ape and child. *Monographs of the society for research in child development*, pages i–252, 1993.

[225] E. S. Savage-Rumbaugh, S. G. S. S. Savage-Rumbaugh, T. J. Taylor, S. Shanker, et al. *Apes, language, and the human mind*. Oxford University Press on Demand, 1998.

[226] S. Savage-Rumbaugh, W. M. Fields, P. Segerdahl, and D. Rumbaugh. Culture prefigures cognition in Pan/Homo bonobos. *Theoria. Revista de Teoría, Historia y Fundamentos de la Ciencia*, 20(3):311–328, 2005.

[227] S. Savage-Rumbaugh, W. M. Fields, and J. P. Taglialatela. Language, speech, tools and writing. A cultural imperative. *Journal of Consciousness Studies*, 8(5-6):273–292, 2001.

[228] S. Savage-Rumbaugh, P. Segerdahl, and W. M. Fields. Individual Differences in Language Competencies in Apes Resulting From Unique Rearing Conditions Imposed by Different First Epistemologies. In *Emory Symposia in Cognition, Oct, 2002, Atlanta, GA, US*. Lawrence Erlbaum Associates Publishers, 2005.

[229] S. Save-Rambaugh. First InterSpecies Musical Conversation. `https://www.youtube.com/watch?v=jg7TaUrXOC8`, 2001. Online; accessed 06-June-2021.

[230] P. Sawers. The Social Dilemma: How digital platforms pose an existential threat to society. `https://venturebeat.com/2020/09/02/the-social-dilemma-how-digital-platforms-pose-an-existential-threat-to-society/`, 2020. VentureBeat; accessed 02-November-2020.

[231] K. W. Scangos, G. S. Makhoul, L. P. Sugrue, E. F. Chang, and A. D. Krystal. State-dependent responses to intracranial brain stimulation in a patient with depression. *Nature Medicine*, pages 1–3, 2021.

[232] C. Schein and K. Gray. The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, 22(1):32–70, 2018.

[233] J. Seymour and P. Tully. Weaponizing data science for social engineering: Automated E2E spear phishing on Twitter. *Black Hat USA*, 37:1–39, 2016.

[234] A. Siani and S. A. Marley. Impact of the recreational use of virtual reality on physical and mental wellbeing during the Covid-19 lockdown. *Health and Technology*, 11(2):425–435, 2021.

[235] F. Silva, R. Ferreira, A. Castro, P. Pinto, and J. Ramos. Experiments on Gamification with Virtual and Augmented Reality for Practical Application Learning. In *International Conference in Methodologies and intelligent Systems for Techhnology Enhanced Learning*, pages 175–184. Springer, 2021.

[236] P. Singer. Speciesism and moral status. *Metaphilosophy*, 40(3-4):567–581, 2009.

[237] M. Slater, S. Neyret, T. Johnston, G. Iruretagoyena, M. Á. de la Campa Crespo, M. Alabèrnia-Segura, B. Spanlang, and G. Feixas. An experimental study of a virtual reality counselling paradigm using embodied self-dialogue. *Scientific reports*, 9(1):1–13, 2019.

[238] J. Sliwa, A. Planté, J.-R. Duhamel, and S. Wirth. Independent neuronal representation of facial and vocal identity in the monkey hippocampus and inferotemporal cortex. *Cerebral cortex*, 26(3):950–966, 2016.

[239] R. Smith, K. Friston, and C. Whyte. A Step-by-Step Tutorial on Active Inference and its Application to Empirical Data. *PsyArXiv*, 2021.

[240] N. Spatola and K. Urbanska. God-like robots: the semantic overlap between representation of divine and artificial entities. *AI & SOCIETY*, 35(2):329–341, 2020.

[241] C. Stanford. *Significant others: The ape-human continuum and the quest for human nature*. Basic Books, 2001.

[242] L. Stern. What Can Bonobos Teach Us About the Nature of Language? `https://www.smithsonianmag.com/science-nature/bonobos-teach-humans-about-nature-language-180975191/`, 2020. Online; accessed 03-June-2021.

[243] I. Stevens and F. Gilbert. N-of-1 Trials for Closed-Loop Deep Brain Stimulation Devices. *Ethics & human research*, 42(2):28–33, 2020.

[244] E. Sue Rumbaugh, I. Roffman, E. Pugh, and D. M. Rumbaugh. Ethical methods of investigation with Pan/Homo bonobos and chimpanzees. In *Biocommunication: Sign-Mediated Interactions between Cells and Organisms*, pages 449–573. World Scientific, 2017.

[245] Y. Sullivan, M. de Bourmont, and M. Dunaway. Appraisals of harms and injustice trigger an eerie feeling that decreases trust in artificial intelligence systems. *Annals of Operations Research*, pages 1–24, 2020.

[246] M. Sumitani, M. Osumi, H. Abe, K. Azuma, R. Tsuchida, and M. Sumitani. A Robot Has a Mind of Its Own Because We Intuitively Share It. *Applied Sciences*, 10(18):6531, 2020.

[247] J. P. Taglialatela, S. Savage-Rumbaugh, and L. A. Baker. Vocal production by a language-competent Pan paniscus. *International Journal of Primatology*, 24(1):1–17, 2003.

[248] A. Tallón-Ballesteros. Exploring the Potential of GPT-2 for Generating Fake Reviews of Research Papers. *Fuzzy Systems and Data Mining VI: Proceedings of FSDM 2020*, 331:390, 2020.

[249] 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.

[250] A. Theodorou and V. Dignum. Towards ethical and socio-legal governance in AI. *Nature Machine Intelligence*, 2(1):10–12, 2020.

[251] C. Thierry. A multimodal corpus of Human-Human and Human-Robot conversations including synchronized behavioral and neurophysiological recordings. In *Late-breaking Track at the SIGDIAL Special Session on Physically Situated Dialogue (RoboDIAL-20)*, 2020.

[252] L. L. Tifft and D. Sullivan. A needs-based, social harms definition of crime. *What is crime*, pages 179–206, 2001.

[253] M. Tomasello. The role of roles in uniquely human cognition and sociality. *Journal for the Theory of Social Behaviour*, 50(1):2–19, 2020.

[254] S. Tombs. For pragmatism and politics: Crime, social harm and zemiology. In *Zemiology*, pages 11–31. Springer, 2018.

[255] N. Toth, K. D. Schick, E. S. Savage-Rumbaugh, R. A. Sevcik, and D. M. Rumbaugh. Pan the tool-maker: investigations into the stone tool-making and tool-using capabilities of a bonobo (Pan paniscus). *Journal of Archaeological Science*, 20(1):81–91, 1993.

[256] R. Tucciarelli, N. Vehar, and M. Tsakiris. On the realness of people who do not exist: the social processing of artificial faces. *PsyArXiv*, 2020.

[257] 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.

[258] M. P. Van Den Heuvel and O. Sporns. Rich-club organization of the human connectome. *Journal of Neuroscience*, 31(44):15775–15786, 2011.

[259] W. Van der Wagen and W. Pieters. From cybercrime to cyborg crime: Botnets as hybrid criminal actor-networks. *British journal of criminology*, 55(3):578–595, 2015.

[260] R. Van Noorden. Publishers withdraw more than 120 gibberish papers. *Nature News*, 2014.

[261] T. Vinnakota. A cybernetics paradigms framework for cyberspace: Key lens to cybersecurity. In *2013 IEEE International Conference on Computational Intelligence and Cybernetics (CYBERNETICSCOM)*, pages 85–91. IEEE, 2013.

[262] M. von Heimendahl, R. P. Rao, and M. Brecht. Weak and nondiscriminative responses to conspecifics in the rat hippocampus. *Journal of Neuroscience*, 32(6):2129–2141, 2012.

[263] L. S. Vygotsky. *Mind in society: The development of higher psychological processes*. Harvard university press, 1980.

[264] L. S. Vygotsky. Thinking and speech. *The collected works of LS Vygotsky*, 1:39–285, 1987.

[265] J. P. Wahle, T. Ruas, N. Meuschke, and B. Gipp. Are neural language models good plagiarists? A benchmark for neural paraphrase detection. *arXiv preprint arXiv:2103.12450*, 2021.

[266] D. M. Wegner and K. Gray. *The mind club: Who thinks, what feels, and why it matters*. Penguin, 2017.

[267] B. Wernaart. Developing a roadmap for the moral programming of smart technology. *Technology in Society*, 64:101466, 2021.

[268] Wikipedia. Blockchain. https://en.wikipedia.org/wiki/Blockchain, 2021. Online; accessed 05-May-2021.

[269] K. Williford, D. Bennequin, K. Friston, and D. Rudrauf. The projective consciousness model and phenomenal selfhood. *Frontiers in Psychology*, 9:2571, 2018.

[270] V. A. Wilson, C. Kade, S. Moeller, S. Treue, I. Kagan, and J. Fischer. Macaque gaze responses to the primatar: a virtual macaque head for social cognition research. *Frontiers in Psychology*, 11:1645, 2020.

[271] S. Witteveen and M. Andrews. Paraphrasing with large language models. *arXiv preprint arXiv:1911.09661*, 2019.

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

[273] M. A. Wood. Rethinking how technologies harm. *The British Journal of Criminology*, 2020.

[274] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, and A. K. Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17(2):151–178, 2020.

[275] H. Y. Yan, K.-C. Yang, F. Menczer, and J. Shanahan. Asymmetrical perceptions of partisan political bots. *New Media & Society*, page 1461444820942744, 2020.

[276] S. Zafari and S. T. Koeszegi. Attitudes toward attributed agency: role of perceived control. *International Journal of Social Robotics*, pages 1–10, 2020.

[277] S. Zeadally, E. Adi, Z. Baig, and I. A. Khan. Harnessing artificial intelligence capabilities to improve cybersecurity. *IEEE Access*, 8:23817–23837, 2020.

[278] L. Zhang and V. L. Thing. Three Decades of Deception Techniques in Active Cyber Defense-Retrospect and Outlook. *Computers & Security*, page 102288, 2021.

[279] 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.

[280] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.

[281] D. Zügner and S. Günnemann. Adversarial attacks on graph neural networks via meta learning. *arXiv preprint arXiv:1902.08412*, 2019.


9 789090 352008