Cyborgnetic Analyses

Epistemic Security Augmentation – *The Homo Cyborgneticus Metamorphosis*

Dr. Nadisha-Marie Aliman, M.Sc.

Epistemic Security Augmentation – *The Homo Cyborgneticus Metamorphosis*

Epistemische veiligheidsverbetering

– *de homo cyborgneticus metamorfose*

(met een samenvatting in het Nederlands)

Contents

Chapter 1

Introduction

Epistemic security [463] is related to the protection of a society's knowledge creation and knowledge communication processes. In the present information ecosystem permeated by colloquial uses of expressions such as "post-truth" [87], "fake news" [324] and "deepfakes" [457], epistemic threats can manifest themselves in a variety of ways including e.g. nefarious attention dynamics [273, 463], epistemic stagnation linked to "filter bubble" mechanisms [487], the erosion of trust [87], but importantly also intentional epistemic distortion conducted by malicious adversaries [462] which can encompass the misuse of technology such as AI [113]. The latter engenders a hardened collective agreement on empirical observations (which we term automated disconcertion) caused by the mere possibility of deceptive epistemic artefacts such as misleading deepfakes[1]. Against this backdrop, one can state that the use of present-day AI already permeated the epistemic infrastructure at an international level. In short, while present-day AI can give rise to tremendously beneficial effects for society, it can simultaneously increase the severity of epistemic security issues [119] with regard to scope, speed and scale. In this vein, the main twofold research question of this *transdisciplinary* book can be formulated as follows: 1) how could one *mitigate AI-related epistemic security risks* in the deepfake era and 2) which strategies could support a responsible *epistemically-sensitive AI design* that is informed of 1)? Given the complexity of the underlying socio-psycho-techno-physical epistemic threat landscape, this book coalesces knowledge from various domains such as e.g. cybersecurity-oriented AI safety, psychology, cybernetics, virtual reality (VR), human-computer interaction, philosophy, natural language processing and creativity research.

Past work proposed countermeasures to combat AI-related epistemic threats including maliciously crafted deepfakes and "fake news" more broadly. For instance, the technical detection of AI-generated content [46, 524, 569] has been thematized. A study [212]

[1]Here, the term "deepakes" refers to any deep-learning based technology harnessed for the generation of synthetic artefacts agnostic of the intentionality and thus also encompassing beneficial use cases [149].

mentioned governmental restrictions on AI hardware and training data next to "proof of personhood" schemes and another suggested the development of "truthful AI" [172]. 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*" [77] and that *"it is essential to foster increased civic literacy of the nature of one's interactions"* [77]. Another analysis presented a set of defense measures against the spread of deepfakes [113] 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. Previous work also performed different assessments on the severity of more general epistemic threats in the deepfake era. For instance, concerning "fake science news" and their impacts on *"credibility and reputation of the science community"* [252], 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"* [347]. Following the author, scientists could equip citizens with sense-making tools without which *"emotions and beliefs that pander to false certainties become more credible"* [347]. It has been stated that the existence of deepfake videos confronts society with severe epistemic threats [176]. Thereby, it is assumed that *"deepfakes reduce the amount of information that videos carry to viewers"* [176] which analogously quantitatively affected the amount of information in text-based news due to earlier "fake news" phenomena. Beyond that, an "epistemic babble" scenario [462] has been postulated as worst-case scenario in which society looses its ability to distinguish between "truth and fiction" [462]. Mainly, a common theme underlying many rather gloomy narratives referring to concepts such as "post-truth world" [307], "post-epistemic world" [260] and "epistemic anarchy" [280] seems to be an empiricism-based epistemology with the epistemic aim to obtain *justified* and *truer beliefs* via (probabilistic) belief updates given *evidence*. However, as explained in this book, empiricist epistemology is not without any alternative and an epistemic doom is not inevitable.

Generally, both for epistemic security and for epistemically-sensitive AI design, one may first need to improve the epistemic assessment of present-day AI itself. For this, in the first part of the book from Chapter 2 to 7, we explain why epistemic security cautions society against both *over*estimating the epistemic capacity of present-day AI and *under*estimating its yet underexplored facets including especially the augmentation of anthropic creativity. Instead of following an empiricist line of reasoning that would risk to sabotage itself e.g. by unnecessarily causing the mental construction of a post-epistemic stance, we craft epistemic defense strategies taking the epistemological philosophy of *critical rationalism* as developed by Popper [410, 411] and reinvigorated by Frederick [200, 202] as point of departure. In the course of the book, we gradually refine this framework. We discuss why in the long term, instead of habitually focusing on data-based "evidence" and the sources of knowledge – which may appear to be useful short-term heuristics but which

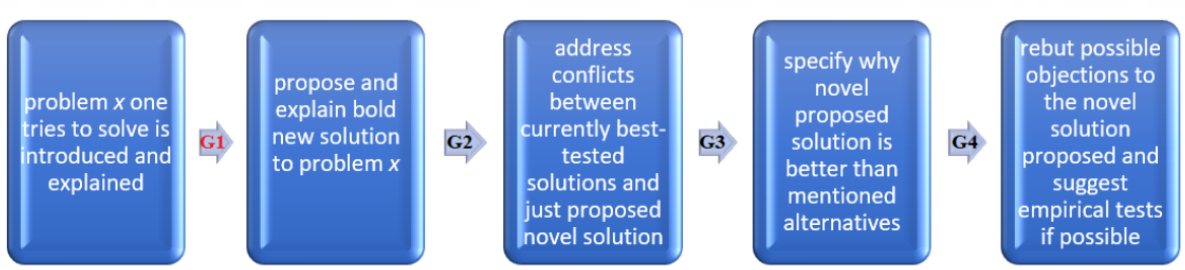

Figure 1.1: Exemplary epistemic total order for the generation of new EBs (the instructions are loosely inspired by an essay of Frederick [201]). Each glue operation x is indicated via a label G_x. EBs are a special form of explanatory information (EI) obtained by interweaving EI blocks via the step-by-step application of rational procedures sampled from a robust explanation-anchored, adversarial and trust-disentangled epistemology. Thereby, "trust-disentangled" signifies that the epistemic modus operandi is grounded in agreed upon criteria for *better* EBs i.e. it is orthogonal to any trust relation between involved entities – which means a better EB must be formulated such that metaphorically speaking it appears to defend itself against adversarial candidate EBs. The inherently comparative criteria for better EBs are updatable and determined by agreement. Current criteria encompass e.g. a preference for explanations that are simpler, provide more novel falisifiable predictions, are more innovative, more aesthetically appealing than rival ones. In science, the specification of (direct or indirect) empirical tests in G_4 is the default condition.

could obviously be forged with more ease in the deepfake era, one must foreground the *content* of knowledge. Thereby, in line with Popperian philosophy, our epistemic aim should be to strive for ever better new *explanations* embedded in a process of *bold* novel conjectures and (provisional) refutations [3, 200, 411]. This book expounds why in the deepfake era, our epistemic aim should be to craft ever better new instances of what we call *explanatory blockchains* (EBs) – a special form of chains of explanations[2] respecting a robust epistemic format. For a simple exemplary illustration on a generic recipe for a new EB, see Figure 1.1. More specifically, we conjecture *that* (and later discuss *why* in the third part of the book) there exists an *information-theoretical asymmetry* between the ability to *create* information of a type x versus the ability to *understand* that information of type x. Consistent with earlier research [14], we define as Type *II* entities all entities for which it is possible to (consciously) *understand* linguistic explanations and as Type *I* entities all entities for which this is impossible[3]. While humans are exemplary Type II

[2]The recent specific concept of EBs facilitates an extension beyond the vaguer and broader term of "explanatory knowledge" which was frequently utilized by Deutsch [158] but may be problematic because one could e.g. state that present-day language AI *is* able to generate outputs that are colloquially widely perceived as "new explanations". However, no language AI has been collectively agreed upon by scientists and philosophers to have been able to reliably generate arbitrary new yet unknown EBs.

[3]This ontology has *no* relation to the metaphor of Kahneman on "System 1" and "System 2" linked to two modes of human brain functioning with the first one being prediction-dominated/automatic and the second one prediction-error dominated/controlled but both modulated by precision weights [266].

entities, a Type I entity could e.g. be the set of all systems that are presently commonly referred to as AI, a chair, thoughts, non-human mammals, ideas, language itself being a primordial technology, stone tools, fishes and so forth. We state that while there may exist no theoretical limitation to the accuracy with which Type I entities such as present-day AIs could create new non-EB-like information, it is *impossible* for Type I entities to reliably create new EBs with arbitrary high accuracy. The latter is a *scientific* claim since, as recommended by Popper [201, 200, 411], it is a bold universal statement which is easily amenable to experimental problematization and it is refutable by a new better theory that would explain a Type-I-shortcut to the reliable creation of new EBs. We discuss implications in real-world environments, conventional social media and VR.

In Chapter 8, the second part of the book compactly introduces a novel paradigm for epistemically-sensitive AI design termed the *Conjecture, Observe, Orient, Co-Create, Act* (COOCA) loop. We focus on the generic concept of a *cyborgnet* [16], a template for a dynamic, hierarchical and context-dependent functional unit that can be described by a directed graph where explanatory narratives combine *at least* one Type *II* entity with *at least* one Type *I* entity. (A *cyborgnet* is a highly generic term that is *not* to be confused with the much more narrow concept of a cyborg. Since the cyborgnetic approach generically regards tools including language as a form of technology, the first language-cognizant humans already instantiated a cyborgnet. Thus, both an individual early human at the dawn of language and a modern cyborg equipped with an eyeborg such as Neil Harbisson [285] are an example of a cyborgnet. Moreover, multiple humans can act as one cyborgnet and it is possible to construct higher hierarchies of cyborgnet networks including complex nested variants.) We explain why in high-risk contexts, in order to achieve a *meaningful* control of present-day AI, one must instantiate a COOCA-loop where *each single* of the five COOCA functions is cyborgnetic and should not allow an OODA-loop [14] constellation if one or more function(s) form a Type-*I-only*-pipeline.

In the third part of the book, in Chapter 9, we conduct a short comparative analysis to explicate *why* there exists a *qualitative* epistemic gap between present-day AI and humans – which instantiates a so-called *cyborgnetic comprehension bottleneck*. Synthesizing modern transdisciplinary knowledge from i.a. systems theory, biology, neuroscience, philosophy of creativity and physics we provide novel explanations for that phenomenon. The book elucidates why both for epistemic security and epistemically-sensitive AI design in the deepfake era, humans may profit from a generic self-reflective epistemic process termed the *homo cyborgneticus metamorphosis*. The latter refers to the procedure of consciously theorizing the so-called *cynet butterfly effect* – a new cyborgnetic version of what is referred to as butterfly effect [28, 444, 518] in the context of frameworks describing complex systems. In line with recent accounts of dynamic creativity [128, 129], the fundamental *unpredictability* of cyborgnetic creativity is emphasized. Indicated implications for attempts to create "artificial superintelligence" [79] from scratch are extended in Chapter 11.3.

In the final discussion in Chapter 10, we comment on how in the present deepfake era, it seems that epistemological philosophy can and should become a much more *scientifically* palpable topic. On the whole, it seems that for epistemic security reasons, humanity may need to empower itself with rational self-knowledge to avoid loosing a sense of agency caused by overestimating the forgery capabilities of present-day AI. An augmented rationality is required to avoid a passive condition steered by meaningless Type-I-loops with even lethal consequences. Thereby, rationality in the deepfake era cannot passively reduce itself to forgeable confirmation-based empiricist strategies and heuristics. Instead, there is the need for a rationality that also promotes the creation of ever better new EBs. Moreover, we expound that the mentioned process of the homo cyborgneticus metamorphosis, while appearing uniquely suited for the modern deepfake era, can as well be understood as a rediscovery of timeless cyborgnetic knowledge. We build a link to early epistemic postulates stemming from *Indian philosophy* [393, 496]. The latter may offer a new perspective on the concept of *self-transcendence* [302] analyzed in modern psychology [290] and positive computing [95, 364] – a potential inspiration for future epistemically-sensitive AI design. Finally, in Chapter 11, we provide a set of novel cyborgnetic epistemic requirements via which one could extend beyond the narrow imitation-based concept of Turing Tests. Overall, because the epistemic strategies presented in this book are developed and refined *gradually*, each chapter from Chapter 2 to 8 ends with an *epistemic meta-analysis* that contextualizes the chapter against the backdrop of the whole book.

On the whole, this book provides the following 7 main contributions:

1. Chapter 2 provides recommendations on how to perform a *trandisciplinary AI observatory* of international scope with relevance for the augmentation of epistemic security. We illustrate the framework with numerous concrete practical examples.

2. In Chapter 3, 4 and 7, we conduct *cybersecurity-oriented design fictions* grounded in threat models for AI-related epistemic security solutions in VR settings.

3. In Chapter 5, we introduce the novel concept of scientific and empirical adversarial AI attacks of which so-called *deepfake science attacks* are a subset.

4. In Chapter 6 and 7, we present new *cyborgnetic creativity augmentation* strategies unifying (epistemic) security and ethics endeavors with language AI as use case.

5. Chapter 8 extends beyond the OODA-loop and introduces the cyborgnetic *COOCA-loop* as new *meta-paradigm* for a responsible *epistemically-sensitive AI design.*

6. In Chapter 9 and 10, we elaborate on the *cynet butterfly effect* and the implications of cyborgnetic epistemology for epistemic security in the deepfake era. We connect the *homo cyborgneticus metamorphosis* to early insights from *Vedantic philosophy.*

7. Chapter 11 contains new ideas for future strict *scientific AI evaluation* frameworks.

Outline

- Chapter 2 introduces an international transdisciplinary AI observatory project with epistemically-relevant recommendations.
- Chapter 3 examines epistemic security in VR and explains why strictly speaking, we do neither inhabit a post-truth era nor a post-falsification era.
- Chapter 4 uses immersive journalism as a use case to elucidate how cybersecurity-oriented immersive design fictions grounded in threat models could be utilized to mitigate AIVR risks that affect epistemic security.
- Chapter 5 introduces scientific and empirical adversarial (SEA) AI attacks using cyber threat intelligence and scientific writing as use cases.
- Chapter 6 collates a set of generic cyborgnetic defenses against SEA AI attacks specifically affecting science and education.
- Chapter 7 thematizes the role of cyborgnetic creativity augmentation and new explanatory blockchains for a VR-based epistemic security training and an epistemically-sensitive threat modelling – both for VR and real world environments.
- Chapter 8 describes the COOCA-loop meta-paradigm.
- Chapter 9 presents the cynet butterfly effect and the hereto linked process of the homo cyborgneticus metamorphosis.
- Chapter 10 concludes, provides an overview of cyborgnetic epistemology, specifies practical recommendations for AI regulation and design and establishes a connection between the homo cyborgneticus metamorphosis and Vedantic philosophy.
- Chapter 11 discusses future research directions attempting to craft better epistemic assessment frameworks of cyborgnetic nature that qualitatively extend beyond imitation-based Turing Tests.

List of Publications

This book is based on the following set of 7 papers enumerated in chronological order:

- N. 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.
- N.-M. Aliman and L. Kester. Facing Immersive "Post-Truth" in AIVR?. *Philosophies*, 5(4), 45, 2020.
- N.-M. Aliman, L. Kester and R. Yampolskiy. Transdisciplinary AI Observatory – Retrospective Analyses and Future-Oriented Contradistinctions. *Philosophies*, 6(1), 6, 2021.
- N. Aliman and L. Kester. Epistemic defenses against scientific and empirical adversarial AI attacks . In *Proceedings of the Workshop on Artificial Intelligence Safety 2021 co-located with the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI 2021), Virtual, August, 2021.*, 2021.
- N.-M. Aliman and L. Kester. Moral Programming: Crafting a flexible heuristic moral meta-model for meaningful AI control in pluralistic societies. *Wageningen Academic Publishers*, (2022): 63-80, 2022.
- N.-M. Aliman and L. Kester. "Immoral Programming: What can be done if malicious actors use language AI to launch 'deepfake science attacks'?". *Wageningen Academic Publishers*, (2022): 179-200, 2022.
- N. Aliman and L. Kester. VR, Deepfakes and Epistemic Security. In *2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 93-98. IEEE, 2022.

Chapter 2

Transdisciplinary AI Observatory

This chapter is based on a modified form of the publication: N.-M. Aliman, L. Kester and R. Yampolskiy. Transdisciplinary AI Observatory – Retrospective Analyses and Future-Oriented Contradistinctions. *Philosophies*, 6(1), 6, 2021. As the first author of the underlying paper, I had a vital contribution. It was solely my responsibility to write down the content and to perform an extensive literature research and in-depth analysis.

2.1 Motivation

Lately, the importance of addressing AI safety, AI ethics and AI governance issues has been acknowledged at an international level across diverse AI research subfields [29, 146, 175, 193, 269, 506]. From the heterogeneous and steadily growing set of proposed solutions and guidelines to tackle these challenges, one can extract an important recent motif, namely the concept of an *AI observatory* for regulatory and feedback purposes. Notable early practical realizations with diverse focuses include Italian [493], Czech [309], German [352] and OECD-level [385] AI observatory endeavors. Thereby, the Italian AI observatory project targets the public reception of AI technology and the Czech one tackles legal, ethical and regulatory aspects within a participatory and collective framework. The German AI observatory jointly covers technological foresight, administration-related issues, sociotechnical elements and social debates at a supranational and international level. Finally, the OECD AI Policy Observatory *"aims to help policymakers implement the AI Principles"* [385] that have been pre-determined by the OECD and pertain among others to data use and analytical tools. Theoretical and practical recommendations to integrate the retrospective documentation of internationally occurring AI failures have been presented by Yampolskiy [550] and very recently McGregor [356]. In addition, Aliman [14] proposed to complement such *reactive* AI observatory documentation efforts with *transdisciplinary* and *taxonomy-based* tools as well as *proactive* security activities.

In this chapter, we build on the approaches of both Yampolskiy and Aliman and elaborate on the necessity of a *transdisciplinary AI observatory* integrating both reactive and proactive retrospective analyses. As reactive analysis, we propose a taxonomy-based *retrospective descriptive analysis* (RDA) which analytically documents factually already instantiated AI risks. As proactive analysis, we propose a taxonomy-based so-called *retrospective counterfactual risk analysis* [539] (RCRA) that inspects plausible peak *downward counterfactuals* [436] of those instantiated AI risks to craft future policies. Downward counterfactuals pertain to *worse* risk instantiations that *could* have plausibly happened in that specific context *but did not*. While an RDA can represent a suitable tool for a qualitative overview of the current AI safety landscape revealing multiple issues to be addressed in the immediate near-term, an RCRA can supplement an RDA by adding breadth, depth and context-sensitivity to these insights with the potential to improve the efficiency of future-oriented regulatory strategies[1].

The remainder of the chapter is organized as follows. In the next Section 2.2, we first introduce a simple fit-for-purpose AI risk taxonomy as basis for classification within RDAs and RCRAs for AI observatory projects. In Section 2.3 and in the subsequent Section 2.4, we elaborate on aims but also limitations of RDA and RCRA while collating *concrete examples* from practice to clarify the proposed descriptive and counterfactual analyses. In Section 2.5, we exemplify the requirement for *transdisciplinarily* conceived *hybrid cognitive-affective* AI observatory approaches and more generally AI safety frameworks. In Subsection 2.5.1, we provide *near-term* guidelines directly linked to the practical factuals and counterfactuals from RDA and RCRA respectively. Hereinafter, we discuss *differentiated* and bifurcated *long-term* directions through the lens of two recent AI safety paradigms: artificial stupidity (AS) and eternal creativity (EC) – succinct concepts which are introduced in Subsection 2.5.2. We provide incentives for future constructive dialectics by delineating central distinctive themes in AS and EC which (while overlapping with regard to multiple near-term views) exhibit pertinent differences with respect to long-term AI safety strategies. Thereafter, in Section 2.6, we briefly comment on data collection methods for RDAs and idea generation processes for RCRAs. Finally, in Section 2.7, we summarize the introduced ensemble of transdisciplinary and socio-psycho-technological recommendations combining retrospective analyses and future-oriented contradistinctions.

[1]Note that in cyborgnetics [16], a meta-disciplinary approach to the mitigation of socio-psycho-techno-physical harm in cyborgnets, next to an RDA and an RCRA, the set of future-oriented countermeasures projected to the immediate counterfactual future is called *future-oriented counterfactual defense analysis* (FCDA). For this reason, the triadic method of RDA, RDA-based RCRA and RDA-and-RCRA-based FCDA *implicitly* utilized in this transdisciplinary observatory can be understood as an early exemplary instantiation of a cyborgnetic analysis.

How and When did Type I system become Dangerous		Causes	
		On Purpose	By Mistake
Timing	Pre-Deployment	*a*	*c*
	Post-Deployment	*b*	*d*

Figure 2.1: Simplified overview of main Type I AI risks. Modified from [19].

2.2 Simple AI Risk Taxonomy

For simplicity and means of illustration, we utilize the streamlined AI risk taxonomy displayed in Figure 2.1 for the classification of practical examples of AI risk instantiations in the RDA and corresponding downward counterfactuals in the RCRA. This simplified taxonomy has been derived from a recent work by Aliman et al. [19]. (Note that the original taxonomy makes a *substrate-independent* difference between two *disjunct* sets of systems: *Type I* systems and *Type II* systems. While the set of *Type II* systems includes all systems that exhibit the ability to *consciously create and understand explanatory knowledge*, Type I systems are by definition all those systems that *do not* exhibit this capability. Obviously, all present-day AI systems are of *Type I* whereas *Type II* AI is up to now *non-existent*. In fact, the only currently known group of *Type II* systems are human entities. For this reason, the taxonomy we consider here for RDA and RCRA only focuses on the practically-relevant and already instantiated classes of *Type I* AI risks.) Following cybersecurity-oriented approaches to AI safety [14, 84, 407, 550], we do not only classically zoom in on *unintentional* failure modes but also on *intentional* malice exhibited by malevolent actors. This distinction is reflected in the utilized taxonomy by contrasting AI risks brought about by malicious human actors (risk *Ia* and *Ib*) vs. those caused by unintentional failures and events (risks *Ic* and *Id*). Moreover, the taxonomy distinguishes between AI risks forming themselves at the pre-deplyoment stage (*Ia* and *Ic*) vs. those forming themselves at the post-deployment stage (*Ib* and *Id*).

2.3 Retrospective *Descriptive* Analysis (RDA)

2.3.1 Aims and Limitations

To allow for a human-centered AI governance, one requires a dynamic responsive framework that is updatable by design [25] in the light of novel emerging socio-technological [24, 96, 353] AI impacts. For this purpose, it has been postulated to combine proactive and reactive mechanisms in AI governance frameworks in order to achieve an effective socio-

technological feedback-loop [25]. An RDA can be understood as a reactive AI governance and AI safety mechanism. More precisely, taxonomy-based RDA documentation efforts could facilitate a detailed especially qualitative overview and valuable opportunity for fine-grained monitoring of the AI safety landscape. It could be harnessed to guide regulatory efforts, inform policymakers and raise sensitivity in AI security, law and the general public. Further, an RDA could inform future ethical and security-aware AI design and guide endeavors to build defense mechanisms for AI systems enhancing their robustness and performance.

In addition to the proposed fourfold qualitative distinction via the classification in risks *Ia*, *Ib*, *Ic* and *Id*, one could also introduce a quantitative parameter for intensity ratings [461] such as harm intensity [14]. Given the harm-based nature of human cognitive templates in morality [217, 456], a harm parameter could provide a meaningful shortcut to encode the urgency of addressing specific risk instantiations in practice. However, given the simultaneous perceiver-dependency [216, 456] of harm perception in morality which is strongly based on dyadic considerations (the degree to which an intentional agent is *perceived* to inflict damage to a vulnerable patient [456]), corresponding assignments may not generalize. Nevertheless, identifying *peaks* of harm intensity above a certain agreed upon threshold (e.g. starting at the level of lethal risks) from an RDA might represent a responsible strategy with less controversial assignments. (Analogously, as further specified in Section 2.4, it is meaningful to focus on analytically derived above threshold downward counterfactuals as basis for an RCRA.) Extracted RDA peaks can be useful to calibrate regulations where necessary while avoiding superfluous constraints for multiple stakeholders that could hinder freedom and progress in the AI field.

Obviously, the quality of RDA results depends on data collection methods and an RDA may not reveal a comprehensive overall picture. Generally, AI risk instantiations could stay unreported, overlooked by the manual or automated data sampling or even remain unnoticed in certain contexts despite already existing. Finally, it is important to note that an RDA should *not* be understood as means to *predict* the future. As known from Popper, a society cannot predict the contents of its own future knowledge [410]. This *fundamental* unpredictability is directly relevant to understand limitations of an AI observatory – it can only reveal patterns of the past. There is no guarantee of repetitions and for instance completely unknownable novel threats could emerge via future human malevolent creativity in the form of risk instantiations *Ia* and *Ib* or via yet unknown errors leading to future instances *Ic* and *Id*. Instead of conceiving of an RDA as an oracle, we suggest framing it as a valuable preparative but incomplete tool with certain fundamental and further non-fundamental limitations. How an RCRA can be utilized to tackle one restriction of the latter sort is described in Section 2.4.

2.3.2 RDA for AI Risk Instantiations *Ia* and *Ib* – Examples

To clarify the implementation of a taxonomy-based RDA for an AI observatory, we briefly analytically document a variety of concrete already instantiated AI risks starting with those linked to *intentional* malice (AI risks *Ia* and *Ib*). For risk *Ia*, the current goals of the human entities in the context of many induced events are mostly either *adversarial* goals hold by malicious actors or *research* goals of white hats and AI security researchers. To provide a simple and compact overview for risk *Ia*, we group the space of these different goals in a set of 6 (unquestionably non-exhaustive) main clusters: 5 adversarial clusters and 1 research cluster conflating the research goals. The aim of the research cluster is to demonstrate the feasibility of malicious AI design motivated by diverse adversarial goals across a variety of domains in order to foster safety-awareness. Beyond that, we consider 1 extra emerging risk pattern, namely *automated disconcertion* which we introduce in a few paragraphs.

First, an *adversarial cluster 1* could be described as grouping the use of generative AI for subsequent (cyber-)crime facilitation e.g. via impersonation [244, 437, 442, 486]. Striking examples for *adversarial cluster 1* include a deep-learning based voice cloning of the CEO of a UK-based company that enabled a fraudster to acquire ca. $243,000 [486] and a scammer that suceeded to cause a transfer of ca. $287,000 with a deepfake video sample impersonation [437]. Second, one can indentify an *adversarial cluster 2* related to defamation, harassment, revenge and sextortion [210] typically employing deepfake techniques such as deep learning based facial replacement to visually place targeted often female individuals in pornographic video settings they never partook [7]. Third, *adversarial cluster 3* comprises the use of AI for misinformation and disinformation purposes [9] including via fake profiles camouflaged with AI-generated synthetic portraits [431]. Fourth, an *adversarial cluster 4* consists in using deepfake methods (as well as recent applications of deepfakes to virtual reality [121]) for a form of non-consensual voyeurism whereby even underage victims are assumed to be affected in some cases [236]. Fifth, *adversarial cluster 5* includes AI-supported espionage [133] (e.g. via AI-generated fake profile pictures on social media platforms[450]), AI-aided intelligence gathering [416] and controversial AI-supported targeted profiling [365].

Moreover, we identify a *research cluster 1* as described. Notably, security researchers provided proof-of-concepts among others related to designing camouflaged undetectable fake samples usable for other crimes (e.g. adversarial deepfakes bypassing deepfake filters [370] which could be misused to conceal unethical illegal material disguised as deepfakes and furthermore undetected AI-generated fake comments i.a. on a federal public comment website [564]). Recent security work also successfully explored advanced deepfake techniques for improved impersonation, spear-phishing and large-scale disinformation [384]. Yampolskiy crafted a proof-of-concept for an AI-generated fake academic article [278]

perhaps simultaneously acting as cautionary example and as a form of honeypot [566] for inattentive readers that might cite this article unknowingly. Other researchers identified an emerging interest for deepfake ransomware [371] in certain cybercriminal circles. Beyond that, it has been demonstrated that via a replica of a victim intelligent system (a deep reinforcement learning agent), the policies of the victim system can be compromised in a targeted way [107].

Interestingly, an already perceptible consequence of *the mere existence* of risk *Ia* instantiations containing the design of deepfake technologies already led to the emergence of a risk pattern which has been termed automated disconcertion. Automated disconcertion can imply the intentional or also unintentional mislabelling of real samples as fake – e.g. in the context of misleading conspiracy theories [482] or against the background of uncertain political settings as it was the case in Gabon not long ago [234]. (To summarize the latter, a *"recent failed military coup in the context of pre-existing political unrest in Gabon was partially grounded in the proliferation of the wrong assumption that an official presidential video represented a manipulative deepfake video"* (see Chapter 4).) Conversely, automated disconcertion can also mean that fake samples are considered as being authentic or simply lead to highly uncertain and inconclusive settings in which doubts cannot be further resolved in reasonable time with acceptable resources. In short, this additional outlier risk pattern is called *automated* disconcertion since it does not further necessitate the interference of any actors to be repeatedly instantiated after initiation.

Coming to risk *Ib*, its instantiations are currently predominantly concentrated in a single research-oriented cluster (in analogy to *research cluster 1* for risk *Ia* instantiations). However, it is thinkable that exploits of AI vulnerabilities unknown to the public are already taking place before disclosure (a type of zero-day exploits [67] applied to the AI domain). The main benign research goal for security researchers to target risk *Ib* instantiations is currently mostly to disclose existing AI vulnerabilities against malicious attacks and explore possible novel defenses against those before their exploitation. This already led to an incessant attacker-defender race in the fast moving field of security for machine learning and adversarial examples [103, 100, 398, 497]. In recent years, researchers have among others developed different attack schemes on how to evade cybersecurity AI [298], e-mail protection, verification tools [417], forensic classifiers [102] and person detectors [543], how to elicit algorithmic biases [14, 526], how to fool medical AI [111, 192, 231, 570], law enforcement tools [572] as well as autonomous vehicles [97, 413], how to perform denial-of-service and other adversarial attacks on commercial AI services [110, 332, 541], how to cause energy-intense and unnecessarily prolonged processing time [469] and how to poison AI systems post-deployment [117].

2.3.3 RDA for AI Risk Instantiations *Ic* and *Id* – Examples

In this subsection, we continue to elucidate the practical application of a taxonomy-based RDA by now briefly analytically documenting various already instantiated *unintentionally* triggered risks that formed themselves at the pre- and post-deployment stage (i.e. risk *Ic* and *Id* respectively). For risk *Ic*, we group the space of observed failure modes in a set of 5 (unquestionably non-exhaustive) main failure clusters. In addition, we present 1 extra emerging risk pattern. In analogy to the outlier risk pattern of automated disconcertion related to risk *Ia* instantiations, we introduce the risk pattern of *automated peer pressure* representing an already perceptible side-effect of specific risk instantiations *Ic*. In the case of AI risk instantiations *Id*, we consider a single main failure cluster. (Overall, in some cases, it is difficult to delineate a risk instantiation type unambiguously (e.g. *Ic* vs. *Id* in the presence of multiple complex influences or even in a few cases *Ic* vs. *Ia* given different ethical perspectives). This practical limitation is partially linked to the perceiver-dependency of classification-related assignments that may also play a role in a future AI observatory. However, by publicly sharing the sources, it is possible for entities external to an AI observatory to refine interpretations. Generally, we humbly subscribe to the epistemological view that all knowledge is fallible [115].)

For risk *Ic*, we consider the 5 main failure clusters described in the following. First, *failure cluster 1* comprises ethically-relevant instances of algorithmic bias [381]. Part of this cluster are misclassifications of diverse underrepresented patterns in AI training datasets with unethical repercussions as exhibited in e.g. facial misidentification [249], facial recognition failures [91, 144], inaccuracy in AI-aided diagnosis [322]. Other cases are datasets with historically outdated unethical labels [414] and ethically-sensitive training biases favoring overrepresented patterns [272]. Second, *failure cluster 2* refers to instances of poorly designed low-performing AI that are halted subsequently [292]. Third, *failure cluster 3* are AI methods designed for law enforcement but threatening privacy [265]. Fourth, *failure cluster 4* subsumes all unintentional risk instantiations linked to more or less hidden pseudo-scientific or outdated and previously refuted preconceptions. For instance, the deployment of AI for facial recognition of criminals based on *"minute features"* [143, 241] in their face is based on pseudo-scientific assumptions [400]. Further, the deployment of present-day image-based "emotion recognition" AI is not grounded in state-of-the-art [53] affective science and lacks the required multimodal and context-sensitive modelling to be able to mimick how humans *infer* [208] (and not detect) affective patterns. In fact, a ban has been requested for premature emotion AI i.a. to prevent usage in ethically sensitive settings [137] such as law enforcement, fraud detection or recruiting. Fifth, *failure cluster 5* is linked to affective, persuasive [334] and (micro-)targeted AI-aided methods that already permeate human cognitive-affective constructions in a way extending *beyond the initial design purposes* and causing *epistemic* biases ranging from a loss of critical stance via AI-empowered social media [273, 453] to flawed mind perception in present-

day robots [114, 545].

A further risk pattern that emerged via *the mere existence* of specific AI risk instantiations *Ic* assignable to the *failure cluster 5*, is a construct that we call *automated peer pressure*. It is already known that attention at a collective level can be *intentionally* biased and manipulated in social media [273] also with the help of bots [390, 415] (risk *Ia*). Moreover, as stated in an open letter written by multiple known psychologists and sent to the American Psychological Association: *"[...] the desire for social acceptance and the fear of social rejection are exploited by psychologists and other behavior change experts to pull users into social media sites and keep them there for long periods of time"* [329] – especially children [334]. Susceptible collective attention mechanisms and beliefs are already even *unintentionally* [453] strongly influenced by AI-empowered social media initially developed for benign purposes. Paired with the strong social dependency of humans where social pressure plays an important regulatory role with biological roots [494], it already triggered what one could call automated peer pressure, a self-perpetuating pattern of social pressure [30, 43, 198, 228, 485] without the need for social agents that directly and consciously exert it. Beyond that, the known group phenomenon of *"self-reinforcing networks of like-minded users"* [415] encountered in social media has been termed *homophily* [273, 415]. Overall, a combination of a multiplicity of heterogeneous factors of which epistemic biases, homophily, affective contagion [184, 273], bots and automated peer pressure are only a subset may foster the documented spread of propaganda in social media [415] as well as the reported negative impacts on the mental health of young users [342, 453].

Finally, concerning AI risk *Id*, we observe one main failure cluster which is connected to unanticipated post-deployment usage modes and contexts which also includes eventual complications within unusual interactions of the AI system in a dynamically changing environment. Notable examples are failures of facial recognition AI linked to COVID-19 causing the widespread use of facial masks [321, 366, 372], the invariant responses of natural language processing systems when faced with nonsensical instead of usual meaningful queries [310] (disclosing the low level of understanding) and the AI-based censorship of a picture displaying ancient slavery settings due to a forerunning misclassification labelling the sample as displaying nudity [491]. Other cases include unknown latent biases in medical AI [154] and other forms of biases in medical AI that unfold post-deployment as a function of geographical factors [291].

2.4 Retrospective *Counterfactual* Risk Analysis (RCRA)

2.4.1 Aims and Limitations

While *upward counterfactuals* of a factual event refer to the better ways in which that event could have unfolded but did not, *downward counterfactuals* refer to those conceivable ways in which this event could have turned out worse. In the past, counterfactual thinking has often been framed as detrimental rumination or even as cognitive bias. However, a modern explanatory framework from social psychology termed *functional theory of counterfactual thinking* (abbreviated with FTCT in the following) stresses that counterfactual thoughts can offer *"[...] insights that comprise blueprints for future action [...]"* [436]. FTCT stresses that counterfactual thinking serves problem-solving and can exhibit high usefulness especially in *complex multi-causal domains* [436]. At the intrapersonal level, counterfactual thoughts are based on implicit processes caused by problems, they are linked to a negatively valenced state of core affect [171] and have the potential to evoke (mental or physical) actions that can potentially correct the underlying errors. This procedure instantiates a regulatory loop – which corresponds to a type of negative feedback model [171] enacted as goal-oriented corrective behavior.

Recently, the notion of an RCRA [539] building upon downward counterfactuals from historical events has been proposed to risk stakeholders in the context of risk management applied to hazardous events (such as earthquakes or terroristic attacks). As explained by Woo [539], such an innovative augmented historical analysis represents a generic universal tool that can supplement regulatory resilience tests and sense-making while facilitating the formation of more differentiated and nuanced views. Given its conjectured *domain-general* nature and seeming applicability to complex multi-causal domains of risk analysis, we suggest to transfer RCRA to AI observatory contexts *at a conceptual level*.

For illustration purposes, the Subsection 2.4.3 presents a simplified *RDA-based RCRA* which directly builds upon the exemplary RDA performed in the last Section 2.3. Our method is *loosely* inspired by Woo's RCRA conception which manifests itself by the general integration of downward counterfactuals from historical samples. However, our stepwise methodology (elucidated in the subsequent Subsection 2.4.2) to extract meaningful candidates for the simulation[2] of downward counterfactuals given a large state space of past events has been independently conceived and tailored to the specific AI observatory domain. Overall, we understand an RCRA as complement for a forerunning RDA. Together, this pair of retrospective analyses could provide a solid starting point for future AI observatory projects to be however necessarily updated and error-corrected with time.

[2]Downward counterfactuals can be (co-)created e.g. in a predominantly mental form, facilitated by immersive design fiction settings (including storytelling narratives and virtual reality) or simulated and visualized with technological tools.

In abstract terms, combining RDA and RCRA can be seen as a socio-technological enactment of the regulatory loop-governing behavior [171] described in FTCT – which fits to the AI governance recommendation mentioned earlier in Subsection 2.3.1, namely the notion of a socio-technological feedback-loop combining proactive and reactive measures [14, 24, 25]. While an RDA mainly represents a *reactive* documenting approach, an RCRA *attempts* to broaden future *proactive* measures by anticipating potential extreme branches of the future while resisting the fallacy to cast itself as oracle tool. We emphasize that in the light of the fundamental unpredictability of future knowledge creation as well as the fallibility and incompleteness of human knowledge, surprises and errors are unavoidable. No RCRA can guarantee unassailability. This is similarly the case in cybersecurity for other types of techniques that are likewise assignable to a broad class of proactive security measures related to downward counterfactuals such as penetration testing [531] and red teaming [423, 428]. Also there it holds that the non-detection of a vulnerability does not guarantee its absence. (Conversely, the detection of a vulnerability does also not guarantee its future exploitation by malicious actors[3].)

2.4.2 Preparatory Procedure

After having expounded on aims and limitations of an RDA-based RCRA, we speak to the preparatory procedure of meaningfully extracting the required downward counterfactuals for an RCRA taking as input the set O_{RDA} containing *all instances* from the forerunning RDA. However, before providing further details, we recall as mentioned in Subsection 2.3.1 that a meaningful agreed upon threshold τ of harm intensity is recommendable. Although perceiver-dependent, a sufficiently high threshold such as e.g. when set to plausible downward counterfactuals of *at least lethal* dimension may be suitable. On an oversimplified harm intensity scale with *1* standing for almost no harm and *5* for existential risk, let *4* stand for a lethal risk (with *2* encoding minor and *3* major harm). Naturally, this threshold and scale are solely employed for purely illustrative purposes and more differentiated and tailored approaches may be required in practice [14]. Equipped with the scale and the exemplary threshold $\tau = 4$, we elaborate in the following on how the set O_{RCRA} of all *clusters*[4] considered in an RCRA can be constructed starting with O_{RDA} and consecutively applying the following ordered sequence of 4 operations in yet to be described ways: 1) *taxonomization*, 2) *analytical clustering*, 3) *brute-force deliberation and threshold-based pruning*, 4) *assembly*.

As first step, *taxonomization* is applied to O_{RDA} which consists in a one-to-one mapping

[3]For instance, their interest could shift, the asset could be(come) less interesting or the attack too time-consuming and costly.

[4]For an enhanced context-sensitivity and to avoid overfitting to the idiosyncrasies of single isolated events, we recommend RCRA simulations at the level of clusters and not of single instances as becomes apparent in the next Subsection 2.4.3.

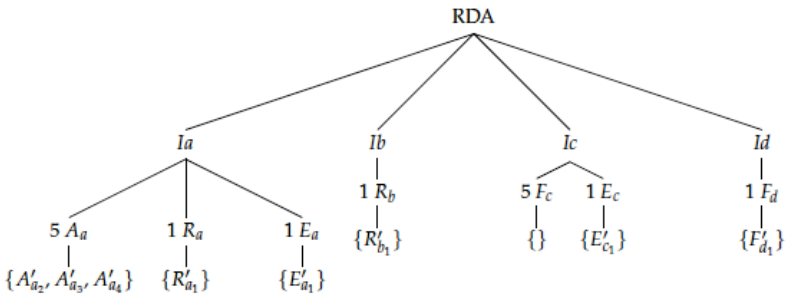

Figure 2.2: Simplified sketch on possible preparatory procedure to extract peak generic downward counterfactuals for an RCRA out of a forerunning taxonomy-based RDA for an AI observatory. The top node stands for the initial set O_{RDA} containing all RDA samples. For illustration, the risk instantiation clusters from Section 2.3.2 and 2.3.3 are filled in. A refers to *adversarial*, R to *research*, E to *extra* and F to *failure* cluster. The conjunction of all analytically derived leaves are possible generic above threshold downward counterfactuals of interest for the RCRA. In this example, the output set for the RCRA corresponds to $O_{RCRA} = \{A'_{a_2}, A'_{a_3}, A'_{a_4}, R'_{a_1}, E'_{a_1}, R'_{b_1}, E'_{c_1}, F'_{d_1}\}$. For more details, see text.

of each AI risk instantiation sample to either a key from the taxonomy (i.e. *Ia*, *Ib*, *Ic* or *Id*) or in theory to a generic placeholder key for novel unknown patterns. In our description, all samples were directly or at least secondarily assignable to the pre-existing taxonomy keys and no unknown key was required. As second step, the researchers apply an *analytical clustering* operation based on a self-generated explanatory semantic grouping linking every sample associated with a specific key, to a cluster. By way of example, under risk *Ia* discussed in Section 2.3.2, this operation led to 5 adversarial clusters, 1 research cluster and 1 extra cluster. In a third step, the researchers apply *brute-force deliberation*[5] *and threshold-based pruning* by mentally going through every single sample of O_{RDA} and trying to devise – within reasonable self-determined time limits – a plausible downward counterfactual where it holds for the self-rated harm intensity h, that $h \geq \tau$. If such a suitable downward counterfactual is generated in time, the sample is *maintained*, otherwise the sample is discarded from further consideration. Finally, the fourth operation *assembly* is performed which requires to assemble O_{RCRA} by linking back the remaining samples to their clusters from the second step. On this basis, one obtains the generic downward counterfactuals that need to be analyzed for the intended RCRA. In short, this simple step-wise procedure takes RDA instances as inputs and produces a set of generic RCRA clusters as output. This output set O_{RCRA} represents the superset of

[5]In theory, this search can be optimized further. However, the aim is to (at a later stage) obtain *a broad as possible* set of counterfactual instances to increase illustrative power. Both one-to-one and many-to-one mappings between downward counterfactual instances and clusters can potentially become RCRA-relevant if stored. This is connected to the *complementary cognitive co-creation* method used to interlink the preparatory procedure with the RCRA that we explain in Subsection 2.6.2.

the searched meaningful generic above threshold downward counterfactuals of interest. For clarification, the next paragraph briefly comments on the application of this simple preparatory procedure to our exemplary RDA instances.

While applying the third step of brute-force deliberation and threshold-based pruning, we deleted a large amount of RDA samples since many instances did *not* seem to have had a plausibe downward counterfactual with a harm intensity $h \geq \tau$. However, we decisively already identified certain rare samples where this condition was fulfilled. In the fourth step, we assembled O_{RCRA} by linking these maintained samples back to 8 RDA clusters as described in the following. For risk *Ia*, we deleted the first and fifth adversarial cluster but maintained *adversarial cluster 2* (encoded with A_{a_2}), *adversarial cluster 3* (A_{a_3}), *adversarial cluster 4* (A_{a_4}), the single *research cluster* (R_{a_1}) and the extra cluster (E_{a_1}) of *automated disconcertion*. For risk *Ib*, the standalone research cluster (R_{b_1}) was maintained. For risk *Ic*, only the extra cluster (E_{c_1}) of *automated peer pressure* remained, and we deleted all failure clusters. Finally, for risk *Id*, we kept the single available failure cluster (F_{d_1}). While these clusters were mapped to *factual* risk instantiations, an RCRA obviously requires the generation of corresponding downward counterfactuals. Thus, instead of A_{a_2}, we encode its unreal generic downward counterfactual which we denote A'_{a_2}. Similarly, instead of A_{a_3}, we write A'_{a_3} and so forth. Consequently, as illustrated in a highly simplified form in Figure 2.2, one can hereafter fairly straightforwardly assemble these fragments (visualized as the leaves of the tree) in order to obtain the final output set $O_{RCRA} = \{A'_{a_2}, A'_{a_3}, A'_{a_4}, R'_{a_1}, E'_{a_1}, R'_{b_1}, E'_{c_1}, F'_{d_1}\}$.

2.4.3 Exemplary RDA-based RCRA for AI Observatory Projects

Recently, co-creation *design fictions* (DFs) [6, 406] known from human-computer interaction (HCI) [426] have been recommended for security practices in the AI field [262] and at the intersection between AI and virtual reality (AIVR). Generally, DFs *"can be used for technological future projections by experts in the form of e.g. narratives or construed prototypes that can be represented in text, audio or video formats but also in VR environments"* (see Chapter 4 where we cautioned likewise *not* to regard a DF as a means to *predict* the future but as preparatory tool). In our view, one promising way to perform an RDA-based RCRA could be to frame each RCRA cluster as co-creation DF task. Distinctively, instead of projecting into the future as performed in classical DF contexts, such an *RDA-based RCRA-DF* construes instances of RCRA cluster narratives (or experiential prototypes) projecting to the counterfactual past. For illustration, we apply a simplified RDA-based RCRA-DF to each of the 8 elements within the set $O_{RCRA} = \{A'_{a_2}, A'_{a_3}, A'_{a_4}, R'_{a_1}, E'_{a_1}, R'_{b_1}, E'_{c_1}, F'_{d_1}\}$ assembled in the preparatory procedure of the previous Subsection 2.4.2. For RCRA-DFs pertaining to intentional malice (risks *Ia* and *Ib*), we provide short DF narratives taking the form of a succinct *threat model* [101]

specifying adversarial goals, knowledge and capabilities. By contrast, for unintentional failure modes (risks *Ic* and *Id*), we instead describe a short failure model comprising initial design goals, knowledge gaps and unintended effects. Generally, we only consider instantiations of RCRA clusters that correspond to above threshold downward counterfactuals (i.e. with a harm intensity $h \geq \tau$ whereby in Subsection 2.4.2, τ was exemplarily set to lethal risk dimensions).

Downward Counterfactual DF Narrative A'_{a_2}

- ***Adversarial Goals:*** AI-aided defamation, revenge, harassment and sextortion.
- ***Adversarial Knowledge:*** Since it is a malicious stakeholder that is designing the AI, the system is available to this adversary in a transparent *white-box* setting. Concerning the knowledge pertaining to the human target, a *grey-box* setting is assumed. 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.
- ***Adversarial Capabilities:*** In the following, we briefly speak to exemplary plausible counterfactuals of at least lethal nature that malicious actors could have been capable to bring about and that are *"worse than what actually happened"* [539] (as per RDA). For defamation purposes, it would have been for instance possible to craft AI-generated fake samples that wrongly incriminate victims with *not* actually executed actions (e.g. a fake homicide but also fake police violence) leading to a subsequent assassination when deployed in precarious milieus with high criminality. To enact revenge with lethal consequences in socio-cultural settings that particularly penalize the violation of restrictive moral principles, similar AI-based methods could have been applicable (e.g. via deepfakes assumingly displaying fake adultery or contents linked to homosexuality). An already instantiated form of AI-enabled harassment mentioned in the RDA consists in sharing fake AI-generated video samples of pornographic nature via social media channels [7]. Consequences could include suicide of vulnerable targets (as generally in cybervictimization [276]) or exposure to a lynch mob. In fact, the contemplation of suicide by deepfake pornography targets has already been reported lately [210]. Finally, concerning AI-supported sextortion, warnings directed to teenagers and pertaining to the convergence of deepfakes and sextortion have been formulated recently [141]. Given the link between sextortion and suicide associated with motifs such as i.a. hopelessness, humiliation and shame [374], consequences of technically feasible but not yet instantiated deepfake sextortion scams could also include suicide – next to simplifying this criminal enactment by adding automatable elements.

Downward Counterfactual DF Narrative A'_{a_3}

- ***Adversarial Goals:*** AI-aided misinformation and disinformation.
- ***Adversarial Knowledge:*** Identical to adversarial knowledge indicated in 2.4.3.
- ***Adversarial Capabilities:*** Technically speaking, a malicious actor could have crafted misleading and disconcerting fake AI-generated material that could be interpreted by extreme endorsers of pre-existing misguided conspiracy theories as providing evidence for their beliefs inciting them to subsequent lethal violence. A historical precedent of gun violence as reaction to fake news seemingly confirming false conspiracy theories was the Pizzagate shooting case where a young man fired a rifle in a pizzeria *"[...] wrongly believing he was saving children trapped in a sex-slave ring"* [225]. Beyond that, when it comes to (micro-)targeted [273] disinformation, conceivable malicious actors could have more systematically already employed hazardous AI-aided information warfare [415] techniques in social media. This could have been supported by AI-enabled psychographic targeting tools [273] and via networks of automated bots [64, 415] partially concealed via AI-generated artefacts such as fake profile pictures. While the level of sophistication of many present-day social bots is limited [35], more sophisticated bots emulating a breadth of human online behavior patterns are already developed [77, 558] and it is known for some time [467] that *"[...] political bots exacerbate political polarization"* [556]. By AI-aided microtargeting of specific groups of people that are ready to carry out violent acts, malicious actors could have caused more political unrest with major lethal outcomes. In fact, Tim Kendall who was a prior director of monetization at Facebook recently stated more broadly that *"[...] one possible near-term effect of online platforms' manipulative and polarizing nature could be civil war"* [453].

Downward Counterfactual DF Narrative A'_{a_4}

- ***Adversarial Goals:*** AI-aided non-consensual voyeurism.
- ***Adversarial Knowledge:*** Identical to adversarial knowledge indicated in 2.4.3.
- ***Adversarial Capabilities:*** Before delving into downward counterfactuals that corresponding malicious actors could have already brought about, it is important to note that the goal considered in this cluster is not primarily the credibility or appearance of authenticity exhibited by the synthetic AI-generated contents. Rather, the focus when visually displaying the target non-consensually in compromising settings is more on feeding personal fantasies or facilitating a demonstration of power [121, 236] while the synthetic samples can obviously concurrently be shared via social media channels. Against this backdrop, it is not difficult to imagine

that when editing visual material of vulnerable targets with practices such as deep-learning based "undressing" [236], a disclosure could induce motifs of hopelessness, humiliation and shame in some of those individuals provoking suicidal attempts similar to the hypothetical deepfake sextortion counterfactual described in 2.4.3. The mere sensing of having been victimized via non-consensual deepfake pornography has also been associated with the perception of a "digital rape" [179, 210]. Especially when the victims are underage [236], this could plausibly reinforce suicidal ideation. Another dangerous avenue may be subtle *combination possibilities* available to the malicious actor. Non-consensual voyeuristic (but also more generally abusive) illegal but quasi-untraceable material bypassing content filters could be meticulously concealed with deepfake technologies and unnoticedly propagated[6] for some time. This could hinder criminal prosecution and particularly threaten the life of vulnerable young victims. Potentiated with automated disconcertion, it could cause a set of latent lethal socio-psycho-technological risks.

Downward Counterfactual DF Narrative R'_{a_1}

- ***Adversarial Goals:*** Research on malevolent AI.
- ***Adversarial Knowledge:*** Identical to adversarial knowledge indicated in 2.4.3.
- ***Adversarial Capabilities:*** To begin with, note that in this RCRA cluster, we assume that the research is motivated by *malign* intentions contrary to the corresponding factual RDA research cluster that is conducted with benign and precautionary intentions by security researchers and white hats as mentioned in Subsection 2.3.2. This additional distinction is permissible due to its property as *downward* counterfactual. By way of illustration, malicious actors could have already performed research on malevolent AI design in the domain of autonomous mobility or in the military domain. They could have developed a novel type of meta-level physical adversarial attacks on intelligent systems[7] directly utilizing other physically deployed intelligent systems under their control. Such an attacker-controlled intelligent system could be employed as a new advanced form of present-day physical adversarial examples [109, 167, 303, 369, 529, 543] against a selected victim intelligent system. The maliciously crafted AI could have been designed to optimize on physically

[6]For instance by mixing real material with synthetic elements obtained from style-based generative adversarial network methods [283], deep-learning based face-replacement and *adversarial deepfake* techniques [370] in order to evade content filters critical to law enforcement.

[7]With intelligent systems, we refer to technically feasible AIs implemented with the intention to let those independently perform an OODA-loop (i.e. observe, orient, decide, act) that is goal-governed by human entities (e.g. using updatable *human-defined* ethical goal functions [25, 14] prepared pre-deployment). Why in high-risk contexts one must however instantiate the so-called COOCA-loop meta-paradigm instead is discussed in Chapter 8.

fooling the victim AI system once deployed in the environment e.g. via physical manipulations at the sensor level such as to misleadingly bring about victim policies with lethal consequences entirely unintended by the operators of the victim model. A further concerning instance of malign research could have been secretive or closed-source research on automated medical AI forgery tools that add imperceptible adversarial perturbations to inputs such as to cause tailored customizable misclassifications. While the vulnerability of medical AI to adversarial attacks is already known [111, 192, 231, 420, 570] and could be exploited by actors intending medical fraud e.g. for financial gains, certain exertions of this practice in the wrong settings could be misused as tool for murder attempts and targeted homicides.

Downward Counterfactual DF Narrative E'_{a_1}

- ***Adversarial Goals:*** This extra cluster of automated disconcertion refers to a risk pattern that emerged automatically from the mere availability and proliferation of deepfake methods in recent years. However, it is conceivable that this AI-related agentless automatic pattern can be intentionally instrumentalized in the service of other (not necessarily AI-related) primary adversarial goals. One example for a primary adversarial goal cluster in the light of which it is appealing for a malicious actor to strategically harness automated disconcertion, would be *information warfare and agitation on social media*. In fact, early cases may already occur [482].
- ***Adversarial Knowledge:*** Identical to adversarial knowledge indicated in 2.4.3.
- ***Adversarial Capabilities:*** The use of social media in information warfare has been described to be linked to the objective to intentionally blur the lines between fact and fiction [273]. The motif of automated disconcertion itself could be weaponized and misleadingly framed as providing evidence for post-truth narratives offering an ideal breeding ground for global political adversaries performing information warfare via disinformation. Malicious actors could then intensify this framing with the use of pertinent AI technology enlarging their adversarial capabilities as described earlier under the cluster of AI-aided misinformation and disinformation in 2.4.3. Given that automated disconcertion may aggravate pre-existing global strategically maintained confusions [118], it becomes clear that a more effective incitement to lethal violence, political unrest with major lethal outcomes or civil wars could be achieved.

Downward Counterfactual DF Narrative R'_{b_1}

- ***Adversarial Goals:*** Research on vulnerabilities of deployed AI systems.

- ***Adversarial Knowledge:*** *Grey-box* setting (partial knowledge of AI implementation details).

- ***Adversarial Capabilities:*** As analogously described in 2.4.3, we assume that the research is conducted with malicious intentions. Zero-day exploits of vulnerabilities in (semi-)autonomous mobility and cooperative driving settings to trigger extensive fatal road accidents seem realizable.

Downward Counterfactual DF Narrative E'_{c_1}

- ***Designer Goals:*** Although automated peer pressure refers to an agentless self-perpetuating mechanism that emerged through AI-empowered (micro-)targeting[8] on social media, its origins can certainly be traced back to the original benign or neutral economic intentions underlying the early design of social media platforms. Psychologist Richard Freed called present-day social media an *"attention economy"* [334] and it is plausible that social media profits from the maximization of utilization time spent by their users.

- ***Knowledge Gaps:*** Early social media designers may not have foreseen the far-reaching consequences of the designed socio-technological artefacts including threats of lethal dimension or even existential caliber according to some present-day viewpoints [453].

- ***Unintended Failures:*** The more attention users pay to social media contents, the more time they may spend with like-minded individuals (consistent with homophily[9] [273, 415]) and the more they may be prone to automated peer pressure. The latter can an also be partially fueled by social bots aggravating polarization [556]. The bigger the success of information warfare and targeted disinformation on social media and the higher the performance of the AI technology empowering it, the more groups of like-minded peers could (but of course not necessarily) uptake

[8] As for instance successfully performed in the Cambridge Analytica case [273].

[9] Homophily in social media is a multidimensional construct that can refer to attitudes, beliefs, preferences, appearances across a variety of domains. It is by no means limited to the often discussed case of political homophily [123]. For example, empirical social media studies identified weight-based homophily [301], journalistic homophily [233], homophily in rumor sharing [323], higher perceived homophily by users from collectivistic cultures [328], perceived homophily driving consumer purchase intentions [448] and credibility of information [270], homophilic effects in consumer-website relationships [296], homophily as factor for vlogger popularity [317], ideological hashtag homophily in marketing campaigns [544] and even homophily related to music preferences [573]. Apart from that, it is known in social psychology that *"ingroups are seen as more variable than outgroups"* [523] (especially in individualistic cultures). This could arguably strengthen the (wrong) perception of engaging in heterogeneous online spaces. However, some studies actually found social media patterns diverging from homophily [40]. Hence, it is important to further assess the context-sensitive nature of the phenomenon in future work.

misleading ideas. Individuals could then – via these repercussions – sense *a social pressure to suppress their critical thinking* and get accustomed to simply copy in-group narratives irrespective of their contents. This scenario could in turn play into the hands of malicious actors of the type mentioned in 2.4.3 and raise the amount and intensity of the lethal and catastrophic scenarios of the sort described in 2.4.3.

Downward Counterfactual DF Narrative F'_{d_1}

- ***Designer Goals:*** Implementation of high-performance AI.
- ***Knowledge Gaps:*** Designers cannot predict the emergence of yet unknown global risks for which no scientific explanatory framework exists (otherwise that would contradict the fundamental unpredictability of future knowledge creation mentioned in Subsection 2.4.1). Given that the past does not contain data patterns of yet never instantiated hazards, the datasets utilized to train "high-performance" AI cannot already have these eventualities reflected in their metrics.
- ***Unintended Failures:*** Exemplary failures that resulted from this unavoidable type of knowledge gap, are multiple post-COVID AI performance issues [320, 420, 475]. Simultaneously, humanity relies more and more on medical AI systems. Would humans have been confronted with a more aggressive type of yet unknown biological hazard requiring even faster reactivity, it is conceivable that under the wrong constellations, the AI systems optimizing on metrics pertaining to the then deprecated old or on the novel but yet too scarce and thus biased datasets [475] could have led to unreliable policies up to the potential of a major risk.

2.5 Discussion

2.5.1 Hybrid Cognitive-Affective AI Observatory – Transdisciplinary Integration and Guidelines

In this Subsection 2.5.1, we compile near-term AI safety guidelines with respect to: 1) the factual RDA clusters introduced in Section 2.3 and 2) the RDA-based RCRA clusters from Subsection 2.4.3. For 2), we only specify the necessary supplementary and non-overlapping guidelines to avoid repetitions.

Near-term Guidelines for Risks *Ia* and *Ib*

RDA:

- A_{a_1}: Clearly, for risk *Ia* instances of *adversarial cluster 1* related to the misuse of generative AI to facilitate cybercrimes (e.g. via impersonation within social engineering phone calls), already known security measures regarding identity check are needed as minimum requirement. A standard approach to mitigate dangers of malevolent impersonation [548] is to go beyond something you are (biometric) [554], and to also require something you know (password) [547] and/or something you have (ID card). Generally, an awareness-raising training of users and employees on social engineering methods including the novel combination possibilities emerging from malicious generative AI design seems indispensable. In addition, it may be helpful to systematically complement those measures with old-fashioned but potentially effective pre-approved but updatable private arrangements made *offline* which can also employ offline elements for identity check. For instance, the malicious actor may not be able to react appropriately in real-time if presented with a from his perspective semantically unintelligible inspection question making use of offline pre-agreed upon (dynamically updated) linguistically ciphered insider idioms. The induced confusion could consequently help to dismantle the AI-aided impersonation attempt. Having said this, it is important to analyze the attack surface that the availability of voice cloning and even video impersonation with generative AI brings about when instrumentalized for attacks against widespread voice-based or video-based authentication methods.

- A_{a_2}: This cluster pertaining to AI-aided defamation, harassment, revenge and sextortion exhibits the need for far-reaching legislatures for the protection of potential victims. Legal frameworks but also social media platforms may need to counteract large-scale propagation of material that threatens the safety of targeted entities. Social services could initiate emergency call hotlines for dangerous deepfake victimization. Moreover, the creation of (virtual or physical) local temporary shelters or havens for affected individuals combining a team of transdisciplinary experts and volunteers for acute phases immediately succeeding the release of compromising material on social media channels appears recommendable. However, the initiation of a societal-level debate and education could foster destigmatization of deepfake instrumentalized for defamation, harassment and revenge. It could dampen the effects of widely distributed compromising material once the general public looses interest in such currently salient elements. More broadly, educating the public about the capabilities of deep-fake technology could be helpful in mitigating defamation, harassment and sextortion since just like society learned to deal with fake Photoshop images, society can also learn scepticism towards AI-generated content.

- A_{a_3}: AI-aided misinformation and disinformation represents a highly complex socio-psycho-technological threat landscape that needs to be addressed at multiple levels using multi-layered [537] approaches. For instance, in a recent work addressing

the malicious applications of generative AI and corresponding defenses, Boneh et al. [77] provide a list of directly or indirectly concerned actors: *"authors of fake content; authors of applications used to create fake content; owners of platforms that host fake content software; educators who train engineers in sensitive technologies; manufacturers and authors who create platforms and applications for capturing content (e.g., cameras); owners of data repositories used to train generators; unwitting persons depicted in fake content such as images or deepfakes; platforms that host and/or distribute fake content; audiences who encounter fake content; journalists who report on fake content; and so on"*. Crucially, as further specified by the authors, *"a precise threat model capturing the goal and capabilities of actors relevant to the system being analyzed is the first step towards principled defenses"* [77]. In fact, as briefly adumbrated in Subsection 2.4.3, the format of the RDA-based RCRA-DFs we proposed for risk *Ia* and *Ib* was purposefully instantiating exactly that – a threat model. Overall, we thus recommend grounding the development of near-term AI safety defenses (as applied to AI-aided disinformation but also more generally) in RDA-based RCRA-DFs that can be once generated potentially retroactively diversified by novel DF narrative instances tailored to the exemplary actors mentioned by Boneh and collaborators. This could broaden the RCRA results and allow for an enhanced targeted development of countermeasures.

- A_{a_4}: For this AI-aided form of non-consensual voyeurism, the measures of an emergency hotline and a specialized haven as mentioned under cluster A_{a_2} are likewise applicable. Legislators need to be informed on psychological consequences especially for underage victims. While cluster A_{a_2} implied the overt public dissemination of compromising material by what minor individuals would be less at risk given the potential repercussions, the purely voyeuristic case can often be *covert* and attracts motivational profiles that can target minor individuals [236]. In addition, it might be valuable to proactively inform the general public and also adult population groups susceptible to this issue in order to lift the underlying taboos and to mitigate negative psychological impacts. In the long run, instantiations of this cluster are unlikely to be prevented any more than one can prevent someone fantasizing about someone else. Hence, in the age of fake generative AI artefacts with the virtualization of fake acts of heterogeneous nature normally violating physical integrity in the real-world, it might become fundamentally important to re-assess and/or update societal notions intimately linked to virtual, physical and hybrid body perception in a critical and open dialogue.

- A_{a_5}: With regard to AI-aided espionage, companies and public organizations in sensitive domains need to broadly create awareness especially related to the risk of fake accounts with fake but real appearing profile pictures. For instance, since the generator in a generative adversarial network (GAN) [214] is by design imitat-

ing features from a given distribution, advanced results of a successful procedure could appear ordinary and more typical – potentially facilitating a psychologically-relevant intrinsic camouflaging effect. In effect, according to a recent study focused on the human perception of GAN pictures displaying faces of fake individuals that do not exist, *"[...] GAN faces were more likely to be perceived as real than Real faces"*[10] [504]. Beyond that, the authors described an increased social conformity towards faces perceived as real independently of their actual realness. This is concerning also in the light of the extra cluster E_{c_1} of automated peer pressure that could make AI-aided espionage easier. A generic trivial but often underestimated guideline that may also apply to AI-aided open-source intelligence gathering would be to reduce the sharing of valuable information assets via social media channels and more generally on publically available sources to a minimum. Finally, to confuse person-tracking algorithms and prevent AI-aided surveillance misused for espionage, camouflage [560] and adversarial patches [543] embedded in clothes and accessoires can be utilized.

- R_{a_1}: As deep-fake technology proliferates and is used in numerous criminal domains, it is conceivable that an arms-race between malevolent fakers and AI forensic experts [38, 458] will ensue, with no permanent winner. Given that this cluster R_{a_1} covers a wide variety of research domains in which security researchers and white hats attempt to preemptively emulate malicious AI design activities to foster safety awareness, a consequential recommendation appears to actively support such research at multiple scales of governance. Talent in this adversarial field would need to be attracted by tailored incentives and should not be limited to a standard sampling from average sought-after skill profiles in companies, universities and public organizations of high social reputation. This may also help to avoid an undesirable drift to adversaries for instance at the level of information operations risking reinforcing capacities mentioned in the downward counterfactual DF narratives on cluster A'_{a_3}, R'_{a_1} and E'_{a_1} presented in Subsection 2.4.3. Hence, a monolithic approach in AI governance with a narrow focus on ethics and unintentional ethical failures is insufficient [14]. Finally, we briefly address guidelines related to a specific R_{a_1} issue concerning *science* (as asset of invaluable importance for a democratic society [439]) that did not yet gain attention in AI safety and AI governance but that makes further inspections appear imperative in the near-term. Namely, targeted studies on *AI-aided deception in science* to produce AI-generated text disseminated as *fake research articles* (see the research prototype developed by Yampolskiy [278] in another research context) and possibly AI-generated audiovisual or other material meant to

[10]Note that on the long-term this could in theory skew the unconscious internal model internet users exposed to more and more synthetic faces have of how human faces look like. Outliers from the real distribution could be met with more surprise at the subpersonal level. However, the latter might already be the case today with the widespread use of enhancing filters on social-media.

display *fake experiments* or also *fake historical samples* (see the recent MIT deepfake demonstration [361] developed for educative purposes). However, this technical research direction requires a supplementation by transdisciplinary experts addressing the socio-psycho-technological impacts and particularly the *epistemic* impacts of corresponding future risk instantiations. We suggest that for a safety-relevant sense-making, AI governance may even need to stimulate debates and exchanges on the very epistemological grounding of science – *before* e.g. future texts written by maliciously designed sophisticated AI bots (also called sophisbots [77]) infiltrate the scientific enterprise with submissions that go undetected. For instance, there is a *fundamental discrepancy*[11] between how Bayesian and empiricist epistemology would analyze this risk vs. how Popperian critical rationalist epistemology would view the same risk. Disentangling this epistemic issue is of high importance for AI safety and beyond as becomes apparent in the guidelines linked to the next cluster E_{a_1} below.

- E_{a_1}: Near-term guidelines to directly tackle this extra cluster associated to automated disconcertion seem daunting to formulate. However, as a first small step, one could focus on how to avoid exacerbating it. One reason why this cluster may seem difficult to address is due to its deep and far-reaching *epistemic* implications pertaining to the nature of falsification, verification, fakery and (hyper-)reality [59] itself. With regard to this feature of epistemic relevance, E_{a_1} exhibits a commonality with the just introduced different risk of AI-aided deception in science. We postulate that in the light of pre-existing fragile circumstances in the scientific enterprise including the emergence of modern "fake science" [259] patterns but also the mentioned fundamental discrepancies across epistemically-relevant scientific stances, AI-aided deception in science could have direct repercussions on automated disconcertion. First, it could for instance unnecessarily aggravate automated disconcertion phenomena in the general public as e.g. the belief in *epistemic threats* [176] could increase people's subjective uncertainty. Second, a reinforced automated disconcertion can subsequently be weaponized and instrumentalized by malicious actors with lethal consequences as generally depicted under the downward counterfactual DF

[11] Bayesian and empiricist epistemological stances placing the *empirical collection of evidence* and the identification of *true beliefs* at the center of science may link AI-aided deception to *"epistemic threats"* [176] – knowledge-relevant impairments of belief-updating which they already see emerging via deepfakes (i.a. subsuming a general decrease of information in audiovisual samples [176]). By contrast, Popperian epistemic views [411] and especially their Deutschian extension [158] predominantly emphasize in the first place the *explanatory* and *criticism-centered* purpose of science next to the (experimental) *falsifiability* of hypotheses. Strikingly, Deutsch describes science as the endless quest for invariant, *hard-to-vary explanations* of reality [158]. On this view, AI-aided deception in science may be practically problematic, but without question solvable. In fact, while the empiricist direction faces epistemic threats and a post-truth difficulty, the Popperian and Deutschian direction may neither see explanatory knowledge, truth, falsifiability nor the scientific method per se at risk.

narrative E'_{a_1} described in Subsection 2.4.3. This explains our near-term AI governance recommendation to address AI-aided deception in science as transdisciplinary collaborative endeavor analyzing socio-psycho-technological and epistemic impacts.

- R_{b_1}: For this cluster linked to risk *Ib* and pertaining to research on AI vulnerabilities currently performed by security researchers and white hats, we recommend (as analogously already explained in R_{a_1}) to recruit such researchers preemptively. In this vein, Aliman [14] proposes to *"organize a digital security playground where "AI white hats" engage in adversarial attacks against AI architectures and share their findings in an open-source manner"*. For the specific domain of intelligent systems, it is advisable to proactively equip these AIs with technical self-assessment and self-management capabilities[12] [24] allowing for better real-time adaptability for the eventuality of attack scenarios known from past incidents or proof-of-concept use cases studied by security researchers and white hats. However, it is important to keep in mind that challenges from this cluster also deal with zero-day AI exploits, they are the unknown unknowns and cannot be meaningfully anticipated and prevented, though it is realized that many issues could be caused by under-specification in machine learning systems [148].

RCRA (additional non-overlapping guidelines):

- A'_{a_2}: Generally, one possible way to systematically reflect upon defense methods for specific RCRA instances (generated from *downward* counterfactual clusters) of harm intensity $h_{down} \geq \tau$, could be to perform corresponding *upward* counterfactual deliberations targeting a harm intensity $h_{up} < \tau$. As briefly introduced in Subsection 2.4.1, upward counterfactuals refer to those ways in which a certain event could have turned out better but did not. Recently, Oughton et al. [391] applied a combination of downward and upward counterfactual stochastic risk analyis to a cyber-physical attack on electricity infrastructure. In short, the difference to the method that we propose is that instead of focusing on *slightly better upward counterfactuals given the factual event* as made sense in the case of Oughton et al., we suggest a threshold-based selection of *below threshold upward counterfactuals given above threshold downward counterfactuals*[13]. For instance, as applied to the present downward counterfactual cluster A'_{a_2} which also included a narrative instance describing

[12]The conjunction of technical self-assessment and self-management has been summarized under the synonymous umbrella terms of *Type I AI self-awareness* [14], *self-awareness functionality* [24] or simply *self-awareness*.

[13]Since as mentioned earlier, lower harm intensity may lead to more perceiver-dependent differences, one does not exactly need to establish which exact intensity, one only needs to know *that* it is a non-lethal upward counterfactual scenario.

suicide attempts with lethal outcomes as a consequence of AI-aided defamation, harassment and revenge, it could simply consist in *deliberations on how to avoid these lethal scenarios*. This could be implemented by deliberating from the perspective of planning a human, hybrid or fully automated AI-based emergency team response with a highly restricted timeframe (e.g. to counteract the domino-effect initiated by the deployment of the deepfake sample on social media). Next to a proactive combination of deepfake detectors and content detectors for blocking purposes that can fail, a reactive automated social network graph analysis AI combined with sentiment analysis tools could be trained to detect large harassment and defamation patterns that if paired with the sharing of audiovisual samples, can prompt a human operator. This individual could then decide to call in social services that in turn proactively contact the target offering support as analogously mentioned under the guidelines for the factual RDA sample A_{a_2}.

- A'_{a_3}: For this downward counterfactual cluster on AI-aided misinformation and disinformation of at least lethal dimensions, we focus on recommendations pertaining to journalism-relevant defenses and bots on social media. Disinformation from fake sources could be counteracted with the use of blockchain-based reputation systems [27] to assess the quality of information sources. Journalists could also entertain a collective blockchain-based repository containing all news-relevant audiovisual deepfake samples whose authenticity has been refuted so far. This tool could be utilized as publically available high-level filter to evade certain techniques of disinformation campaigns. Moreover, the case of hazardous large-scale disinformation supported by sophisticated automated social bots is of high relevance for what one can term *social media AI safety*. Ideally, *tests* for a "bot shield" enabling some bot-free social media spaces could be crafted. However, it is conceivable that at a certain point, AI-based bot detection [138] might become futile. Also, social bots already fool people [139, 556] and many assume that humans will become unable to discern them in the long-term. Nevertheless, it could be worthwhile viewing what one could have done better already with present technological tools (the upward counterfactuals) – which can also include the consideration of divergent unconventional solutions or novel formulations of questions. As stated by Barrett, *"[...] progress in science is often not answering old questions but asking better ones"* [51]. Perhaps, in the future, humans could still devise bot shielding tactics that could attempt to bypass epistemic issues [15] intrinsic to imitation game and Turing Test [507] derivatives where "real" and "fake" become relative.

- A'_{a_4}: To tackle suicidal ideation as a consequence of AI-aided non-consensual voyeurism that enters the awareness of the targeted individual, one may need to extend the countermeasures already mentioned in the factual RDA counterpart A_{a_4} of this cluster (which also included the creation of public awareness and the removal of as-

sociated taboos). Social services and public institutions like universities and schools could offer emergency psychological interventions for the person at risk. Next to necessitated measures at the level of legal frameworks to protect underage victims, the subtle case of adult targets calls for instance for a civil reporting office collaborating with social media platforms which could initiate a critical dialogue with the other party to bring about an immediate deletion or at least categorical refraining from further dissemination of the material which can be calibrated to the expectations of the target. Recently, the malicious design of deepfakes has been described as a *"[...] serious threat to psychological security"* [396]. Adult targets may despite the synthetic nature of the deepfake samples and often eventually their private character restricted to a personal possession of the agent in question, perceive their mere existence as degradation [387] – a phenomenon certainly requiring social discourses in the long-term. For a principled analytical approach, an extensive psychological research program integrating a collaboration with i.a. AI security researchers could be helpful in order to be able to contextualize relevant socio-psycho-technological aspects against the background of advanced technical feasibility. Importantly, instead of limiting this research to deepfake artefacts in the AI field, one needs to also cover novel hybrid combination possibilities available for the design of non-consensual voyeuristic material. Notably, this includes blended applications at the intersection of AI and virtual reality [121] (or augmented reality [344]).

- R'_{a_1}: Concerning proactive measures against future research where an adversary designs self-owned intelligent systems to trigger lethal accidents on victim intelligent systems, one might require legal norms setting minimum requirements on the techniques employed for the cybernetic control of systems deployed in public space. From an adversarial AI perspective, this could include the obligation to integrate regular updates on AI-related security patches developed in collaboration with AI security researchers and white hats that also study advanced physical adversarial attacks. This becomes particularly important as many stakeholders are currently unprepared in this regard [315]. As guideline, we propose that future adversarial AI research endeavors explore attack scenarios where adversarial examples on physically deployed intelligent systems are delivered by another physically deployed intelligent system which potentially offers more degrees of freedom to the malicious actor. From a systems engineering perspective, any intelligent system might need to at least integrate multiple types of sensors and check for inconsistencies at the symbolic level. Next to explainability requirements, a further valuable feature to create accountability in the case of accidents could be a type of self-auditing via self-assessment and self-management [24] allowing for a retrospective counterfactual analysis on what went wrong.

- E'_{a_1}: As its factual counterpart E_{a_1}, this counterfactual cluster E'_{a_1} refering to auto-

mated disconcertion instrumentalized for AI-aided information warfare and agitation on social media with the risk to incite lethal violence at large scales, represents a weighty challenge of international extent. As for E_{a_1}, multi-level piecemeal tactics of constructive small steps such as e.g. targeted methods to avoid exacerbating it may be valuable. Concerning AI governance, that could include the strategies mentioned under E_{a_1} but also more general efforts in line with international frameworks that aim to foster strong institutions and error-correction via life-long learning (see e.g. [25] for an in-depth discussion).

- R'_{b_1}: For this counterfactual cluster pertaining to malicious research on vulnerabilities of deployed AI systems with the goal to trigger extensive fatal road accidents, we recommend tailored measures analogous to those presented for the counterfactual cluster R'_{a_1}.

Near-term Guidelines for Risks *Ic* and *Id*

As can already be realized from the scope of the AI safety guidelines proposed in Subsection 2.5.1 which are grounded in our AI observatory exemplification of RDA and RCRA, modern AI technology cannot be analyzed in isolation. In our view, due to the complex multi-causal socio-psycho-technological interwovenness underlying AI risks and their instantiations, AI safety requires an inherently *transdisciplinary*, *hybrid* and *cognitive-affective* approach [14]. Transdisciplinarity is especially required to avoid cognitive blind spots within AI safety risk analyses and formulations of countermeasures or guidelines. AI safety needs a hybrid perspective to incorporate the intricacies of human-computer interactions necessitating a consideration of *human nature* next to purely technological viewpoints. Finally, a cognitive-affective perspective is called for due to the inseparably affective nature of human cognition [55, 299] whose disregard in AI development can consequently engender significant safety issues by virtue of a lack of requisite variety [13]. While the last Subsection 2.5.1 focused on guidelines concerning the AI risks *Ia* and *Ib* related to intentional malice, this Subsection 2.5.1 is linked to the risks *Ic* and *Id* related to mistakes and unintentional failures which are often of ethically-relevant nature. This specific avenue of research represents a well-studied field at the core of modern AI ethics which recognizes multidisciplinarity, human-centeredness and socio-technical contextualization as important requirements [166]. In the last years, a large multiplicity of heterogeneous AI ethics guidelines have been proposed at an international level [195, 227, 363, 536]. We refer the reader to Jobin et al. [275] for a global overview of internationally proposed AI ethics guidelines which are directly of relevance for the 5 failure clusters (F_{c_1} to F_{c_5}) linked to risk *Ic* from the RDA presented in Subsection 2.3.3. In the following, we focus on the few remaining RDA and RCRA clusters which are not classically in the primary focus of AI ethics.

RDA:

- E_{c_1}: This cluster related to automated peer pressure can be i.a. met by measures raising public awareness on the dangers of the confirmation bias [220, 559] reinforced via AI-empowered social media. However, a possible upward counterfactual on that issue would be to revert negative consequences of automated peer pressure by utilizing it for beneficial purposes. For instance, it is cogitable that automated peer pressure need not represent a threat would it simply perhaps paradoxically socially *reinforce critical thinking* instead of reinforcing tendencies to blindly copy in-group narratives. Ideally, such a peer pressure would reinforce *heterophily* (the antonym of homophily) with regard to various preferences with one notable exception being the critical thinking mode itself. Hence, one interesting future-oriented solution for AI governance may be education and life-long learning [25] conveying critical thinking and criticism as invaluable tools for youth and general public. For instance, critical thinking skills fostered in the Finnish public education system were effective against disinformation operations [273]. In fact, critical thinking, criticism and transformative contrariness may not only represent a strong shield to tackle disinformation or automated disconcertion and its risk potentials (cluster E_{a_1} and E'_{a_1} respectively), but it also represents a crucial momentum for human creativity [14, 501]. Generally, peer pressure is in itself a psychological tool that could be systematically used for good, for example by creating an artificial crowd [553] of peers with all members interested in desirable behaviors such as education, start-ups or effective altruism. A benevolent crowd of peers can then counteract hazardous bubbles on social media.

- F_{d_1}: Concerning AI failures rooted in *unanticipated* and yet unknown post-deployment scenarios, it becomes clear that accuracy and other AI performance measures cannot be understood as conclusive and engraved in stone. A possible proactive measure against post-deployment instantiations of yet unknown AI risks could be the establishment of a generic corrective mechanism. Problems which AI systems experience during its deployment due to differences between training and usage environments can be reduced via increased testing and continued updating and learning stages. On the whole, multiphase deployment, similar to vaccine approval phases, can reduce an overall negative impact on society and increase reliability. Finally, for each safety-critical domain in which AI predictions are involved in the decision-making procedure, one could – irrespective of present-day AI performance – foresee the proactive planification of a human response team in case of sudden expanding anomalies that a sensitized and safety-aware human operator could detect.

RCRA (additional non-overlapping guidelines):

- E'_{c_1}: A twofold guideline for this counterfactual cluster (refering to automated peer pressure with lethal consequences via automated disconcertion instrumentalized for AI-aided disinformation), could be to weaken the influence of social bots by measures described under cluster A'_{a_3} and by transforming automated peer pressure into strong incentives for critical thinking as stated in E_{c_1}.

- F'_{d_1}: Finally, for this cluster of major risk dimension being the counterfactual counterpart of cluster F_{d_1}, we emphasize the importance of an early proactive response team formation in contexts such as for instance medical AI, AI in the financial market, AI-aided cybersecurity and critical intelligent cyberphysical assets. In short, AI systems should by no means be understood to be able to truly operate independently in a given task even if current excellent performance measures seem to suggest so. In the face of unknown and unknowable changes, performance is a moving target which if mistaken as conclusive and static could endanger human lifes.

2.5.2 Long-Term Directions and Future-Oriented Contradistinctions

After having introduced a broad variety of near-term guidelines for future AI observatory endeavors based on the exemplified systematic factual and counterfactual retrospective analyses, we provide a differentiated more *general* outlook on explicitly *long-term* AI safety directions. For this purpose, we select two recent theoretical AI safety paradigms: on the one hand a direction that has been termed *artificial stupidity* (AS) (see [546, 498, 499]) and on the other hand, a direction that we succinctly call *eternal creativity* (EC) stemming from recent work [21, 19, 14]. We emphasize that *the paradigm reflected in this book is EC*. Thereby, EC and AS are by no means postulated to represent the full panoply of nuances and views across the entirety of the young AI safety field. Rather, we select these specific two examples because critical contradistinctions ascertainable via a comparative analysis point to a set of decisive bifurcations which might be of particular interest for the AI safety community due to their potentially *axiomatic* relevance for the future of AI research. While AS and EC coincide in multiple short-term considerations given their common *hybrid cognitive-affective* nature and their emphasis on *cybersecurity-oriented* practices, they fundamentally differ with regard to 3 future-relevant contradistinctions.

We consider the following 3 contradistinctive leitmotifs: 1) *regulatory distinction criterium*, 2) *regulatory enactment* and 3) *substrate management*. First, while AS primarily considers *intelligence* levels for 1), EC ponders the ability to consciously create and understand *explanatory knowledge*. Second, whilst AS foresees deliberate *restrictions* of AI capabilities as tool for 2), EC especially tackles their systematic *enhancement*. Third, while

AS views *substrate-dependent* hardware analyses (next to software considerations) for bounded equalization between humans and AIs as approach to 3), EC aims at unbounded *substrate-independent* functional augmentation. While there exist certainly more possible lines along which one could compare AS and EC, we focus on the mentioned 3 themes due to their urgency and potential to foster constructive dialectics in future theoretical and applied AI (safety) research beyond AI observatory contexts. In Subsection 2.5.2 and 2.5.2, we briefly provide a general introduction followed by a summarization of long-term AI safety guidelines formulated from the perspective of AS and EC respectively as seen through the lens of these 3 contradistinctions.

Paradigm Artificial Stupidity (AS)

One core assumption in the AS paradigm is that an artificial general intelligence (AGI) *"[...] can be made safer by limiting its computing power and memory, or by introducing Artificial Stupidity on certain tasks"* [498]. Thereby, an AI system is understood to be made *artificially stupid* on a certain task if its capabilities are deliberately limited by human designers for the purpose of matching the human performance on that task. One mentioned exemplary domain where such a technique is already applied is in text-to-speech synthesis such as e.g in Google Duplex, an AI for natural conversations over the phone whose implementation included *"[...] the incorporation of speech disfluencies (e.g. "hmm"s and "uh"s)"* [330]. Another example is the context of video games where AI can in principle vastly exceed human performance which is however purposefully restricted in order to allow for a positive human-centered gaming experience. More generally, there are many AI application domains where it is human-desirable to mimic anthropic performance or behavioral patterns for an improved customer service. These cases correspond to a type of imitation game which only succeeds if the AI does not reveal latent super-human capabilities. From that point of view, the AS paradigm conceives of making an AI artificially stupid as being necessary to making it pass a Turing test [499, 498].

Simultaneously, in the last years, AI achieved superhuman-level performance across more and more tasks. Further, it is assumed in AS that *"[...] AI tends to quickly achieve super-human level of performance after having achieved human-level performance*" [499]. Against this background, AS argues distinguishingly that *"[...] by limiting an AI's ability to achieve a task, to better match humans' ability, an AI can be made safer, in the sense that its capabilities will not exceed humans' capabilities by several orders of magnitude"* [499]. In short, AS postulates that *AI ability needs to be upper-bounded by human performance* since it risks to otherwise become *uncontrollable*[14] once it turns into what Bostrom termed a *superintelligence* – an intellect exceeding human cognitive performance

[14]For an in-depth discussion related to AI uncontrollability and unpredictability, see especially [552] and [551] respectively.

across *"[...] virtually all domains of interest"* [79]. Such a hypothetical future artificial superintelligence is believed to not necessarily be value-aligned with humans (while potentially becoming unintelligible to humans due to the gaps in performance), to be capable of insidious betrayal (a scenario termed *treacherous turn* [79]) and to potentially represent a major risk [60] to humanity.

- ***Regulatory distinction criterium:*** In this light, one can extract *intelligence* (or more broadly "performance" or "cogntive performance" across tasks) as the recurring theme of relevance for regulatory AI safety considerations under the AS paradigm. At a first level, one could identify two main safety-relevant clusters: a cluster comprising all AIs that are less or equally capable than an average human [499] and another cluster of superintelligent AI systems. The latter can be further subdivided into three classes of systems as introduced by Bostrom [79]: 1) speed superintelligence, 2) collective superintelligence and 3) quality superintelligence. According to Bostrom, the first ones *"can do all that a human intellect can do, but much faster"*, the second ones are *"composed of a large number of smaller intellects such that the system's overall performance across many very general domains vastly outstrips that of any current cognitive system"* and the third ones are *"at least as fast as a human mind and vastly qualitatively smarter"* [79].

- ***Regulatory enactment:*** In a nutshell, AS recommends *limiting* an AI in hardware and software such that it does not attain any of these enumerated sorts of superintelligence since *"[...] humans could lose control over the AI"* [498]. AS foresees regulatory strategies on *"how to constrain an AGI to be less capable than an average person, or equally capable, while still exhibiting general intelligence"* [499].

- ***Substrate management:*** To limit AI abilities while maintaining functionality, AS proposes multiple practical measures at the hardware and software level. Concerning the former it proposes diverse restrictions especially pertaining to memory, processing, clock speed and computing [499]. With regard to software, it foresees necessary limits on self-improvement as well as measures to avoid treacherous turn scenarios [498]. Another guideline consists in deliberately incorporating known human cognitive biases in the AI system. More precisely, AS postulates that human biases *"can limit the AGI's intelligence and make the AGI fundamentally safer by avoiding behaviors that might harm humans"* [498]. Overall, the substrate management in AS can be categorized as *substrate-dependent* because the artificial substrate is among others specifically tuned to match hardware properties of the human substrate for at most equalization purposes. In summary, AS suscribes to the viewpoint that AI safety aims to *"limit aspects of memory, processing, and speed in ways that align with human capabilities and/or prioritize human welfare, cooperative behavior, and service to humans"* [546] given that AGI *"[...] presents a risk to humanity"* [546].

Paradigm Eternal Creativity (EC)

According to Deutsch, *"the only uniquely significant thing about humans [...] is our ability to create new explanations [...]"* [158]. He further specifies that explanatory knowledge *"gives people a power to transform nature which is ultimately not limited by parochial factors, as all other adaptations are, but only by universal laws"* [158]. Instead of emphasizing levels of intelligence or of performance across a wide set of tasks when analyzing AI safety issues, EC (the paradigm foregrounded in this book) focuses epistemologically on one unique "task": *the ability to consciously create and understand explanatory knowledge.* Thereby, in EC, explanatory knowledge creation also implies the capability to *consciously* understand. Given that core affect is understood as a fundamental property of consciousness [55, 51] and is linked to cognitive-affective counterfactual deliberations [19], this excludes philosophical zombie themes [205]. (In modern embodied and enactive cognition frameworks [51, 83], consciousness is linked to processes of inference for the cybernetic control of a substrate in an environment connected to allostasis [299] (anticipation of needs before they occur [51]) – integrating predictions and error signals from external and internal milieu. It is on such cybernetic control grounds that affective dynamics give rise to the egocentric virtual first-person perspective of the world [440, 538] familiar to humans and lacking in present-day AI.)

Note that EC's focus on consciously creating and understanding explanatory knowledge is by no means an anthropomorphic assumption forced on AI systems. As elucidated in constructor theory [159, 161], a novel explanatory framework in physics, explanatory knowledge creators (of which currently only the human species is known) are brought to the fore in physics in an entirely non-anthropocentric way. To put it very simply, constructor theory focuses on *possible* vs. *impossible* counterfactuals i.e. what *could* happen given physical laws and *why* (instead of predictions based on laws of motion and initial conditions). On contemplating the set of all physical transformations that would be *possible* in the universe i.e. those that *could* happen, one would notice that the size of the very subset containing those transformations that *actually happen* can be strongly influenced by entities able to create and understand knowledge on how to bring them about [158]. This is how explanatory knowledge creation enters *"the cosmic scheme of things"* [158] and this is also why EC prioritizes the conscious understanding and creation of explanatory knowledge via creativity[15] instead of intelligence.

[15]From a psychological and neurocognitive perspective, EC currently views creativity as a tri-partite evolutionary affective construct with varying degrees of *sightedness* [21] instead of a *blind* evolutionary process without a goal akin to biological evolution – as mistakenly assumed by Popper [19, 163]. This is epistemologically relevant because ideas are *not* created by blind trial and error (as variation and selection in biological evolution). Even if novel idea contents are fundamentally unpredictable a priori, idea variation is partially guided by previous experience, the task and contextual cues i.e. there is a non-zero coupling between variation and selection [163]. Creativity itself could have historical roots in *serendipity* and multi-purpose socially *shared doubt* [14] facilitating in theory error-correction but initially

At first sight, given the fundamental unpredictability of future explanatory knowledge, it might seem dangerous for AI safety. Deutsch mentions that *"no good explanation can predict the outcome, or the probability of an outcome, of a phenomenon whose course is going to be significantly affected by the creation of new knowledge"* [158] and further that this fundamental limitation is something that *"when planning for the future, it is vital to come to terms with it"* [158]. EC agrees. EC recently formulated the *AI safety paradox* [19, 14] stating that value alignment and control are conjugate requirements in AI safety. This means that both prevailing ideals cannot be simultaneously fulfilled. EC also states that *"the price of security is eternal creativity"* [14]. So despite the AI safety paradox, a cybersecurity-oriented and risk-centered AI safety is possible – when reframed *"as a discipline which proactively addresses AI risks and reactively responds to occurring instantiations of AI risks"* [14]. In short, AI safety is not condemned, it just needs to come to terms with the compulsion to keep correcting and creating solutions "ad infinitum".

- ***Regulatory distinction criterium:*** EC distinguishes two *substrate-independent* and disjunct sets 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 [16]. Thereby, a subset of Type I systems can be conscious (such as non-human mammals) and requires protection akin to animal rights. Obviously, *all present-day AI systems are of Type I* and *non-conscious*. Type II AI built from scratch is *non*-existent today.

- ***Regulatory enactment:*** In theory, with a Type II AI, *"a mutual value alignment might be achievable via a co- construction of novel values, however, at the cost of its predictability"* [14]. As with all Type II systems (including humans), the future contents of the knowledge they will create are fundamentally unpredictable – irrespective of any intelligence class[17]. In EC, this signifies that: 1) *Type II AI is uncontrollable*[18] *and requires rights on a par with humans*, 2) *Type II AI could*

largely used to maintain traditions.

[16]EC could be stated to apply a constructor-theoretic distinction to AI safety insofar as it applies a possibility-impossibility dichotomy [159] embedded in an explanatory framework to it.

[17]Under EC, superintelligence is as explained not of distinctive interest. It is also viewed as *not* implying profound *qualitative* differences to human baselines. Following Deutsch, it would be *"[...] subject only to limitations of speed or memory capacity, both of which can be equalized by technology"* [81]. EC views human augmentation as valid transformative defense strategy [21].

[18]Importantly, note that Type II AI uncontrollability does *by no means* imply that a Type II AI is necessarily more dangerous than an arbitrarily designed Type I AI. First, it is important to consider that already an advanced *Type I* AI could lead to existential risks for instance when maliciously designed by malevolent human actors to operate *"at a global scale (e.g. affecting global ecological aspects or the financial system)"* [19]. Second, while it is obvious that a Type II AI *could* be highly dangerous, this also holds for humans including adult terrorists threatening international safety. Overall, it seems a prejudice to assume that Type II AIs that would be members of an open society would *inherently* tend to opt for immutable goals of indifference or extreme violence (see e.g. Hall [229] for an in-depth explanation).

engage in a mutual bi-directional value alignment with humans – if it decides so and 3) *it would be unethical to enslave Type II AI.* (Finally, banning Type II AI is a potential loss of requisite variety and does not hinder malicious actors to do so.) By contrast, regarding Type I AI, EC implies that: 4) *Type I AI is controllable*, 5) *Type I AI cannot be fully value-aligned* across all domains of interest for humans due to an insufficient understanding of human morality, 6) *conscious Type I AI is possible and would require animal-like rights* but it is clearly *non-existent* nowadays.

- ***Substrate management:*** To avoid functional biases [432] due to a lack of diversity in information processing, EC opts for a *substrate-independent* functional view. Irrespective of its specific substrate composition, an overall panoply of systems is viewed as *one* unit with diverse functions. Given Type-II-system-defined cognitive-affective goal settings, a systematic function integration can yield complementary synergies. Notably, EC recommends research on substrate-independent functional *artificial creativity augmentation* [21] (artificially augmenting *human* creativity *and* augmenting *artificial* creativity). For instance, active inference could technically *increase* Type I AI exploratory abilities [206, 446]. Besides that, in Subsection 2.6.2, we apply a functional viewpoint to *augment* RCRA DF generation by human Type II systems for AI observatory purposes.

2.6 Materials and Methods

2.6.1 RDA Data Collection

For the collection of RDA samples utilized for illustration purposes in this chapter, we undertook a simple keyword-based web search limited to articles in the period between 2018 and 2020. The main queries (with associated boolean operators) that we considered were: "artificial intelligence","AI", "autonomous", "neural network", "deepfakes", "AI" AND "bias", "AI" AND "failure", "AI" AND "security", "AI" AND "safety", "AI" AND "attack". While many terms are tailored to the type of keys represented in the taxonomy (*Ia*, *Ib*, *Ic*, *Id*) that served as basis for categorization in the RDA as introduced in Section 2.2, we also considered utmost *general* queries such as "artificial intelligence" in order to do justice to the eventuality that we might identify a novel entirely unexpected categorization pattern. With other words, we also foresaw the possibility of not yet encountered anomalies while analyzing the results. As briefly mentioned in Subsection 2.4.2, such a

Those patterns are *possible* choices posing major risks, but *not inherent* properties of Type II systems – the content of whose future novel ideas and related decisions *cannot* be prophesied a priori. In short, there is no meaningful total order of "dangerousness" according to which one can compare Type I and Type II AIs. To put it plainly: *both* the *worst risk* and the *greatest luck* for a Type II system could be a Type II system.

case would have been assigned to a *generic placeholder key for novel unknown patterns*. It would have called for further scrutiny and eventually for a future enlargement of the taxonomy. However, as mentioned in Subsection 2.4.2, we did not yet identify any novelty of this kind in the discussed RDA. Though, at a lower level, we discovered *atypical* instances of the pre-existing key-determined clusters. We tagged this atypicality by refering to corresponding clusters with the attribute "extra" – which was the case for the extra cluster of automated disconcertion linked to risk *Ia* and the extra cluster of automated peer pressure connected to risk *Ic*.

Self-evidently, the underlying search can be performed in a more sophisticated way in future AI observatory projects. First, a broader range of keys and combinations can be strategically devised in the light of RDA and RCRA results from a previous AI observatory iteration. Second, the efforts can be supported by web crawlers [246]. Third, this could be combined with sentiment analysis tools [472] to detect negatively polarized texts of interest for an RDA. Fourth, the creation of novel datasets for text classification [306] could be undertaken for the pre-existing keys of the taxonomy which might however remain insufficient with regard to placeholders for novel patterns. In this vein, we stress the importance of human analysts for a deep semantic understanding requiring explanatory knowledge especially when it comes to the discovery of subtle novel tendencies within superficially similar text sources. Morever, an intense examination of textual material can lead to a further disentanglement of pre-existing clusters – which could even reveal the need for a broader change of the taxonomic keys. In short, a safety-aware responsible RDA data collection pipeline is not entirely automatable and requires human-level understanding by analysts.

2.6.2 Interlinking RDA-based RCRA Pre-processing and RCRA DFs

As elucidated in Subsection 2.4.2, the preparatory procedure generating candidate RCRA clusters based on RDA instances consisted of 4 consecutive steps: 1) *taxonomization*, 2) *analytical clustering*, 3) *brute-force deliberation and threshold-based pruning* and finally 4) *assembly*. Subsequently, these RCRA clusters served as basis to generate RCRA DFs that we exemplified with short RCRA narratives instantiating these clusters as presented in Subsection 2.4.3. However, for the sake of simplicity, the exact methodological approach to *interlink* the preparatory procedure and the RCRA co-creation DF was not previously characterized. In a nutshell, we utilized a method we call *complementary cognitive co-creation* (CCC). While other methods are thinkable, we encourage considering CCC where possible for reasons described in the next paragraphs. Beforehand, we must specify that purposefully, the set of researchers involved in the preparatory procedure of the RCRA and the set of researchers performing the ensuing RCRA DFs were *disjunct*. For clarity,

we refer to the former as *preparatory group* and to the latter as *executive group*. We explain how a complementary collaborative effort between these groups in the form of CCC can increase the *variety and illustrative power* of RCRA DFs.

After applying taxonomization and analytical clustering to the RDA instances, the preparatory group has been described in Subsection 2.4.2 to perform brute-force deliberation and threshold-based pruning. While a brute-force search could appear suboptimal at first sight, we specifically considered this option in order to allow for the preparatory group to potentially be able to retrospectively *diversify* the generation of instances performed by the executive group given the RCRA clusters. This becomes possible, because whilst the preparatory group goes through every single available RDA instance, it attempts to generate an above threshold downward counterfactual that if identified can later turn out to be *utile to store*. In short, when a downward counterfactual is successfully generated for a given RDA sample, the preparatory group can not only maintain the RDA sample, but also store the generated downward counterfactual instance for later RCRA augmentation purposes. Thereby, as briefly specified, generic RCRA *clusters* were used instead of specific instances as inputs for the RCRA DFs to avoid overfitting to the idosyncrasies of unique events and possibly generate a broader variety of DF scenarios. In fact, by solely providing RCRA *clusters* to the executive group at the start of the DFs, we avoid a potentially biased negative influence by the narrow instances of the preparatory group that fulfilled a different primary function (namely the identification of above threshold patterns). To recapitulate, the preparatory procedure can be more precisely re-explained as follows: the preparatory group undergoes all 4 consecutive steps with the crucial additional detail that the brute-force deliberation and threshold-based pruning operation *also* includes *the storing of a successfully generated downward counterfactual instance* for each maintained factual RDA instance. After this pre-DF processing, the preparatory group delivers the RCRA clusters to the executive group which then engages in generating a variety of narratives instances for each obtained cluster. Post-DF, the executive group compares the generated instances with those imagined by the preparatory group pre-DF. All cases that were not yet considered by the executive group[19] but were generated by the preparatory group, are concatenated to the now augmented DFs. Duplicates are ignored.

This overall sequence of steps presents a theoretical collaborative basis for *an augmentation of co-creation DFs* to which we refer to with CCC. A further tool that may improve the efficacy of CCC is to add a functional viewpoint (i.e. related to information processing in a certain context). On closer inspection, it becomes clear why CCC can profit from a functional or/and (neuro-)cognitive [5, 156, 71] diversity of the partaking researchers.

[19]Note that if given an RCRA cluster, the executive group would not succeed in imagining a corresponding instance for a narrative, there is always at least one back-up instantiation – which corresponds to the narrative envisaged by the participatory group pre-DF (whose identification represented the precondition for this cluster to exist in the first place).

Given that in the human cognitive domain, variety is the norm [106] and heterogeneity can provide requisite variety in complex multi-causal dynamic problem domains [432] necessitating collective learning [5] and innovation [108], it makes sense to explore this potential. For instance, while the preparatory group can especially profit from individuals that excel at convergent thinking, the executive group may benefit from divergent thinkers. Pre-DF, the preparatory group needs to map from *one* factual *instance* to *one* counterfactual *instance*. In the DF, the executive group maps from *one* counterfactual *cluster* to *many* counterfactual *instances*. The former requires a horizontal integration at a low level of abstraction while the latter requires a vertical integration from a higher to a lower level of abstraction revealing the potential for *complementary* synergies[20]. A CCC-based approach combining a preparatory group comprising i.a. individuals with a cognitive profile exhibiting strengths in the former and an executive group i.a. sampled from a pool of individuals with strengths in the latter could increase efficiency, variety and illustrative power of RDA-based RCRA co-creation DFs – critical to raise safety-awareness in experts but also in the public.

2.7 Conclusions

Starting with a *cybersecurity-oriented* fit-for-purpose taxonomy of ethical distinction, we introduced and exemplified a *retrospective descriptive analysis* (RDA) for future AI observatory projects. Subsequently, we elucidated how to craft a complementary *retrospective counterfactual risk analysis* (RCRA) based on downward counterfactuals from the previously extracted factual RDA samples. Motivated by recent work on risk management of hazardous events [539] and the *functional theory of counterfactual thinking* [436] from social psychology, we elaborated on why an RDA-based RCRA may be suitable for risk analyses in a complex multi-causal domain such as AI safety. Thereafter, in the light of the ethical sensitivity of AI risk instantiations, we discussed the use of harm intensity ratings for samples of an AI observatory given the perceiver-dependent, harm-based and dyadic nature of human cognitive templates in morality [456]. For illustrative purposes, we suggested a threshold-based approach focusing the RDA-based RCRA on downward counterfactuals of at least *lethal* dimensions. On the one hand, such a high threshold may engender fewer discrepancies in the moral perception being related to harm. On the other hand, it may simultaneously represent a suitable threshold reinforcing *mortality*

[20]For instance, despite possible significant context-dependent [106] hindrances, dyadic mismatches [76] and disabilities, autistic traits are also paired with enhanced convergent thinking [2], detail-rich thinking [397] and higher verbal creativity [284] while attention deficit hyperactivity disorder traits have been linked to enhanced divergent thinking [258, 534] and enhanced originality and flexibility [535]. Systematically combining these two complementary cognitive profiles under a CCC-oriented approach to RDA-based RCRA-DFs for AI observatory feedback-loops could engender benefits.

salience (i.e. the awareness of one's mortality). From the perspective of a relevant socio-psychological theory denoted terror management theory [218, 480], mortality salience – whose elicitation is also conceivable in co-creation design fictions from HCI including virtual reality settings [116] – may be able to foster safety-awareness and cautionary attitudes [116, 468]. Against the backdrop of the RDA samples collected and our targeted RDA-based RCRA, we formulated the need for inherently *transdisciplinary* and *hybrid cognitive-affective* AI observatory and AI safety strategies. As guidelines for future work, we compiled a rich variety of tailored multi-level *near-term* solutions.

Finally, we provided a differentiated general outlook on *long-term* AI safety directions by axiomatically contrasting two disparate recent AI safety paradigms along relevant contradistinctive leitmotifs. More precisely, we contrasted the *artifical stupidity* (AS) paradigm with the *eternal creativity* (EC) paradigm. While AS and EC share a common cybersecurity-oriented and hybrid cognitive-affective stance with regard to multiple near-term AI safety solutions, they differ fundamentally in many future avenues of research. AS offers *intelligence-focused*, *restriction-based* and tailored *substrate-dependent* long-term guidelines. By contrast, long-term EC guidelines bring into focus *conscious explanatory knowledge creation and understanding* and recommend unbounded *functional augmentation* of *substrate-independent* nature. While AS suggests utilizing human cognitive performance as upper bound for AI capabilities to limit hardware and software parameters, EC takes a cybernetic perspective according to which humans need to jointly augment both human and AI functions – e.g. via a doubly ambiguous artificial creativity augmentation research.

In a nutshell, we collated retrospective analyses complemented by future-oriented contradistinctions in order to: 1) apprise future AI observatory projects using *concrete examples* from practice and technically plausible above threshold downward counterfactuals, 2) thematizing possibly decisive bifurcations in future AI (safety) research and 3) pointing out the requirement of a constructive collaborative dialectical approach addressing those. As stated by Popper, *"while differing widely in the various little bits we know, in our infinite ignorance we are all equal"* [411]. Time might tell whether the assumption that *"the price of security is artificial stupidity"* or rather that *"the price of security is eternal creativity"* [14] (or none of those) turns out to practically solve long-term AI safety problems. Either way, explanatory knowledge *co-creation* can heavily influence whether we will succeed in *understanding* how to transform today's vulnerability awareness and mortality salience into the currently known or unknowable *upward* counterfactuals of our counterfactual future.

2.8 Epistemic Meta-Analysis

In the following, we retrospectively provide additional comments on the chapter. Firstly, we extract the key takeaways relevant for AI-related epistemic security issues. Secondly, we summarize key insights for epistemically-sensitive AI design.

2.8.1 Relevance for AI-Related Epistemic Security Strategies

The chapter introduced multiple risk *Ia* clusters linked to intentional malice at the pre-deployment stage of the present-day AI system that could have nefarious effects on the knowledge creation and knowledge communication processes of a society. This includes e.g. the following epistemically-relevant clusters: the use of generative AI for cybercrime facilitation, the misuse of deepfakes for defamation and harassment, AI-based disinformation, AI for non-consensual voyeurism, AI-supported espionage, adversarial deepfakes to fool deepfake detection and also automated disconcertion. While one should not *over*estimate present-day AI since we conjecture that it is impossible for Type I AI to create new yet unknown explanatory blockchains (EBs) with arbitrary high accuracy (see Chapter 1), the misuse of Type-I-AI-generated new *non*-EB-like information and also the use of old already known EBs to tailor malicious strategies could engender severe epistemic threats. Thus, in light of the mentioned risk clusters, we conclude that epistemic threats could also emerge by the *under*estimation of present-day AI in this regard.

2.8.2 Relevance for Epistemically-Sensitive AI Design

Especially the risk *Ib* clusters related to the possibility of adversarial attacks and adversarial examples against deployed AI systems in conjunction with the risk *Id* cluster pertaining to unanticipated post-deployment failure modes are relevant for epistemically-sensitive AI design. The reason being that it cautions humans not to *over*estimate the capacities of present-day AI being of Type I. Overstatements of AI abilities and hype can endanger epistemic security via misguided expectations. The latter can also include a psychological effect that one could call a *honey mind trap* [16](the assignment of agency and/or experience to present-day AIs all of which are however non-conscious). A responsible epistemically-sensitive AI design would abstain from reinforcing such misplaced mental attributions by users and offer more transparency by specifying the limitations of Type I AI. In the next Chapter 3, we explain why "post-truth" narratives in the deepfake era are an overestimation of present-day AI since epistemically speaking, we neither inhabit a post-truth nor a post-falsification era. Concerning ambitions to implement Type *II* AI from scratch, Chapter 9 expounds what it may imply and why its *universal* difficulty is fundamentally *under*estimated in the AS paradigm – leaving an epistemic vulnerability.

Chapter 3

Facing Immersive "Post-Truth" in AIVR?

This chapter is based on a slightly modified form of the publication: N.-M. Aliman and L. Kester. Facing Immersive "Post-Truth" in AIVR?. *Philosophies*, 5(4), 45, 2020. 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.

3.1 Motivation

In the last years, the information ecosystem was permeated by falsehood-related concepts such as fake news [324], deepfakes [457], fake realities [409], digital fakery [178] as well as more globally fake science [259] and *post-truth* [86]. Regarding fakery and truth in extended reality (XR) settings and thus by extension VR, Slater et al. [473] recently argued that: *"Society is based on the premise that sensory experiences give ground truth. XR at societal scales has the capacity to decouple sensory experience from ground truth, potentially undermining some core elements of social fabric."* Moreover, it has been stated that the deployment of AI deepfakes may foster the acquisition of false memories [336]. This could be conceivably exacerbated within future extensions of technically already feasible "VR deepfakes" [78, 121, 344] by the particular aptness of VR to facilitate durable memories [311]. While such issues would already play a role regarding unintentional failure modes elicited by ethically aware actors in AIVR, recent research related to the security and safety of AI [19, 84, 407, 555] and VR [404, 104, 222, 512] respectively emphasizes the need to additionally consider the presence of *unethical* malicious actors. Thereby, to consider intentional malevolent design in AIVR could offer a worst-case scenario analysis [237] that can shed more light on the extent of potential consequences exhibited by

the deployment of AIVR technology but also by simpler cases in AI and VR separately. For instance, when addressing defense methods against AI-generated fakery in future immersive (VR) journalism contexts for disinformation purposes (see Chapter 4), one might gain insights on how to tackle the exposition to non-immersive deepfake artefacts. Simultaneously, it might help to foster the vulnerability awareness of VR users yielding cautionary attitudes towards manipulation.

In this chapter, we focus on immersive falsehood in AIVR (see also Chapter 4), the deliberate construction of fake immersive reality landscapes for malicious ends. Using this example, we contemplate the following question: *"can malicious actors in AIVR exacerbate the presumed post-truth phenomenon via immersive falsehood?"*. Decisively, our answer to this question is that it is the wrong question to ask for various reasons that require to be elucidated. While throughout history, many "rational" traditions were averse to affective motives and attempted to distance themselves from visceral and bodily elements, modern affective science assumes that affect is an inseparable part of cognition and perception [55, 255]. Moreover, VR settings are known for their profound affective impacts on users [89]. Hence, Section 3.2 first elaborates on the epistemological implications of affect as intrinsic ingredient in human cognition and perception – not only in VR. Extending beyond that, Section 3.3 explains why the term "post-truth" may *not* serve as accurate description of the current age. Moreover, VR has been described to offer a rich counterfactual experiential testbed for ethics in technological contexts such as AI [12, 23]. Building on this, Section 3.4 briefly discusses how future affective computing and virtual reality methods could be harnessed for counterfactual and other measures that seek to allow an understanding and debiasing of one's own constructions as a response to immersive falsehood. Finally, Section 3.5 concludes.

3.2 Nested Affective VR Worlds

Before addressing the previously mentioned question related to AIVR and "post-truth", it might be essential to first collate information on the nature of human perception of reality from a transdisciplinary perspective. As famously stated by Feynman [185]: *"Science is a way of trying not to fool yourself."* Against the background of post-truth claims, it seems important to first carefully deconstruct the notion of "ground truth" in different ethically-relevant human contexts:

- **Affective Realism and Social Reality:** As stated by Barrett *"the human brain is anatomically structured so that no decision or action can be free of interoception and affect"* [50]. Thereby, interoception and interoceptive predictions pertain to statistical regularities of the internal milieu of an organism (related to the *body*) [299] while

core affect is seen as a fundamental property of consciousness [51] with especially valence (pleasant/unpleasant) and arousal (activated/deactivated) as crucial components. To put it very simply, according to constructionist theories in psychology, all mental states are based on constructions involving three basic elements [389]: exteroceptive sensory array (related to sensory predictions and information sampled from external world), interoception and prior knowledge including past experience. The hereto linked circumstance that *"affective feelings (incidental or not) naturally infuse our perceptions and give us a sense of confidence that they are valid windows onto the real world"* [540] has been termed *affective realism*. Thereby, human perception imposes cognitive-affective concepts on the world often previously constructed in social reality (abbreviated with SR in the following) and shared via language. In this sense, human perception also exhibits a biologically shaped social nature given that humans reciprocally regulate the biological nervous systems of their social conspecifics[1] [36, 54, 455, 494] via interpersonal physiological dynamics [395] that humans can even remotely bring about using language [50]. Generally, *"human brains are transactive and cannot be considered outside the context of other human brains"* [36]. In our view, affective realism and the embodied nature of cognition are crucial to a further understanding as it stresses that SR is of embodied and *perceiver-dependent* nature [48] – as are mental constructions like emotions [51], moral judgments [216], thoughts, perceptions and so on.

- **Theory-Ladenness:** In science, it is important to separate perceiver-dependent from *perceiver-independent* phenomena which directly pertain to the physical reality (abbreviated with PhyR in the following) that diverse scientific areas attempt to understand. As emphasized by Barrett *"all science relies on human concepts and this is true for the astronomy as it is for the science of emotion"* [49]. For illustrative purposes, Barrett explains that while the existence of celestial bodies in PhyR is perceiver-independent, the status of one celestial body being a planet is not (see the reclassification of Pluto from planet to dwarf planet). In short, humans do not have direct access to the hidden states in PhyR but try to infer those. In this process, one needs to keep in mind that all observations are *theory-laden* which cautions scientists that since one actively samples the environment to gather data, one's prior socio-cultural context, hypotheses and affective predispositions inherently shape what we perceive as information and what as noise. To conclude, even

[1]For instance, social groups reveal an attunement of physiological parameters [454, 395], social relationships act as physiological regulators [187, 454], biobehavorial synchrony serves as scaffold for the maturation of infant brains facilitating social development [36, 181] and the metabolic costs and benefits of interpersonal physiological dynamics modulate social interactions throughout a lifetime [54, 494]. Hence, it is also no surprise that social isolation comes with the physiological burden of less regulatory facilitations [50] and *"lacking social connection qualifies as a risk factor for premature mortality"* [257].

prior to AIVR, our perception of reality was never entirely objective nor did we directly have access to truth which could suddenly get lost by experiencing immersive falsehood. SR is as real as socio-cultural conventions such as language or money. While its embodied constructions contain real physiological ingredients grounded in PhyR, one often tends more to see what one believes than vice versa [204, 312].

- **Nested VR Ground Truth:** An important phenomenological aspect of human experience is its virtual, perspectival and egocentric nature [253, 441] with a simultaneous grounding in PhyR linked to cybernetic control [538]. It has been postulated that human persona inhabit a virtual world generated by the brain [253, 538] and governed by affective dynamics to navigate the physical environment anticipating bodily needs before they occur (this process has also been termed allostasis [132, 299]). More generally, waking time, imagination and dreaming are all assumed to be linked to a virtual reality experience (that we abbreviate with VR_{Mind} in the following) which is generated by the brain for embodied control purposes [538]. In waking time, this virtual experience is directly constrained by PhyR, while dreaming has been described as *"virtual reality proper"* [253] due to the decoupling from external sensory stimulation and blockage of motor actuators (with the exception of e.g. eye muscles). While awake and wearing technical VR headsets and being immersed in a virtual world, a complex novel nested situation occurs, *"a nested form of information flow in which the biological mind and its technological niche influence each other in ways we are just beginning to understand"* [345]. In these scenarios, our VR_{Mind} experience is constrained by both the artificially created VR world and still partially always also PhyR (e.g. simply by having a body and literally sitting, standing or walking during the setting). In short, even without using any VR technology, human experience of the world does not only reflect statistical regularities about PhyR, but consists in goal-directed embodied, affective and theory-laden virtual constructions of a perspectival and perceiver-dependent nature – such as those involved in SR. When using VR, one adds an additional layer of sensory-motor and affective constraints leading to a nested composition. With social VR [70], a novel special case of SR constrained by VR arises and poses new challenges.

3.3 Immersive Falsehood – Post-truth, Post-falsification or Other?

After having analyzed various relevant aspects related to the human perception of reality, one can now re-examine the initial question on whether malicious actors in AIVR could exacerbate assumed "post-truth" phenomena via immersive falsehood.

- ***Post-truth?*** As advanced by Buffachi [86], the perception of a "post-truth" era may be linked to the definition assigned to truth in the first place – especially when truth is associated with *consensus* which seems to be compromised in modern times. We agree with Buffachi to instead utilize the word truth in a much more deflationary manner, namely strictly for scientific endeavors. In our view, consensus is a dominant factor in SR and technologies such as AIVR may be able to profoundly distort features of SR and certain democratic processes. However, when it comes to PhyR, it is obvious that AIVR artefacts do not irreversibly destroy our capability to create refutable conjectures about PhyR. While one could believe that the loss of truth would be exacerbated by AIVR because observations may become unreliable[2], it is important to keep in mind that no repetition of observations can ever provide experimental logically valid justification for a theory [158, 411]. As Karl Popper explained, *induction is logically invalid* and for instance no amount of observed white swans ever proves that all swans are white [411, 447]. He pointed at the asymmetry between falsifiability and verifiability [357] emphasizing falsifiability as one of the most important criteria for scientific theories. While no amount of successful experiments can ever justify a theory i.e. establish its truth, negative experiments can make the theory problematic. (Thereby, note that as elaborated in the Duhem-Quine thesis [239], no experimental falsification attempt can be considered as absolute and conclusive. Consequently, it is the case in practice that only multiple contextualized failures and/or the presence of competitive alternatives contribute to consider the theory as refuted. However, since justifications are logically invalid *on principle* [357, 158, 411], this type of more complex context-aware sophisticated falsificationalism and criticism remains the recommendable alternative.) In short, if one does neither equate truth with social consensus nor scientific truth with justification via observations, immersive falsehood of the future lets the existence of truth untouched – even if not directly accessible. Hence, *there is no reason to assume that humans inhabit a post-truth era.* However, this very asymmetry between falsifiability and verifiability leads to a further complication addressed in the next point.

- ***Post-falsification?*** In our view, a legitimate concern is the ability of malevolent actors in AIVR to compromise material that could be utilized to falsify hypotheses in diverse contexts such as science, history, forensics and journalism with political

[2]In fact, from a Bayesian empiricist point of view which links science to *true beliefs* and *empirical justifications*, deepfakes are already assumed to represent *epistemic threats* [176] gradually emptying audiovisual samples of information. By contrast, Popperian epistemology [411] sees science as an *explanation-based* and *criticism-centered* endeavor with *falsifiability* as decisive criterium – which has been extended by Deutsch [158] who views science as the quest to identify invariant *hard-to-vary explanations* of reality. On that view, deepfakes (and immersive falsehood) do *not* put truth at risk (see Chapter 2 for more details including the safety-relevant urgency to thematize these fundamental Bayesian vs. Popperian epistemic divergences).

repercussions. As stated by Popper, while coherence cannot attest truth, *"inconsistency and incoherence do establish falsehood"* [411]. (However, in Section 3.6 and more extensively in Chapter 5 and Chapter 8, we discuss why one must strictly speaking even improve and update this statement for a stronger grounding of critical rationalism as explained by Frederick [202].) Concerning historical and also forensic sources [268], it is important to analyze whether they exhibit mutual or internal inconsistencies. In other scientific areas, falsification attempts can be more easily repeated, but scientists often rely at least on the honesty of other entities publishing their experimental results (i.e. that other scientists do not deliberately temper their results). For instance, future immersive falsehood in the form of AI-manipulated VR news for disinformation but also defamation and extortion purposes could distort historical and forensic records and exacerbate issues in the information ecosystem. Malicious actors could craft future realistic immersive experiences (e.g. of fake AI-generated confirmatory experiments and research (see Chapter 2)) to undermine the scientific enterprise. With increasing degrees of realism, many scientists may not stay immune against such strategies. At first sight, it might thus seem as if immersive falsehood could compromise falsification (e.g. via future VR deepfakes [78, 121]). Fake memories could be specifically induced in users (see also Chapter 4) that may turn out to be difficult to detect. However, as noted under the last bullet point and known from the Duhem-Quine thesis [239], it is *not* the case that falsification can be experimentally established *in isolation* (mainly due to inherent background assumptions that always play a role). In this vein, it signifies that immersive falsehood would predominantly complicate the falsification process by having the potential to lure humans into wrong background assumptions and slowing down progress. However, while acknowledging these significant impacts of immersive falsehood, this complication seems however to represent *a matter of degree* rather than a matter of kind which is why we postulate that *there is no reason to assume the science-threatening scenario of a post-falsification era.*

3.4 Future Work

In the light of this complex and nuanced landscape related to the worst-case consequences of immersive falsehood, future work could address transdisciplinary countermeasures. For instance, while legal and technical strategies may attempt to harness heuristics to penalize, "detect" and establish accountability for immersive falsehood artefacts which might stay a controversial issue, one needs to anticipate the unavoidable proliferation of at least a part of those within the complex, heterogeneous and dense information ecosystem. Therefore, it may be of importance to additionally develop *reactive* strategies addressing the issue on how individuals can retrospectively deal with the situation of having *already experienced*

samples of immersive falsehood without their knowledge. First, one may for instance need to consciously entertain stronger doubts towards visceral and affective experiences. For this purpose, a real-time affective monitoring [256, 355] e.g. during the consumption of immersive journalism and VR news could be investigated. By way of example, measurable physiological arousal parameters [286, 388, 430] could be visually displayed to the user to encourage a critical stance towards the experienced contents. Second, a type of counterfactual awareness training in VR may promote critical scrutiny by exposing users to design fiction scenarios featuring a conjunction of immersive news samples related to *real* events on the one hand and fake ones based on *plausible counterfactuals* on the other hand (see also Chapter 4). Third, users could experience immersive counterfactual scenarios illustrating the consequences of triggered doubts through AI-generated fakery in immersive or non-immersive news settings and the dangers of false memory uptake. In fact, the *mere existence of deepfakes* has already led to doubts with lethal risk potentials such as e.g. in the context of a failed military coup amidst pre-existing political unrest in Gabon [235, 434]. By making the vulnerability of humans to these sorts of doubts and false memory constructions more palpable, user vigilance might consequently increase in AIVR contexts. This could also be supported via tailored VR experiences successfully elicitating *mortality salience* [116] (i.e. the awareness of one's mortality) – which can motivate *safer* attitudes and behaviors [116, 468, 479]. VR could thus represent a suitable awareness-raising tool for future severe AI(VR) safety risks i.a. by facilitating valuable retrospective counterfactual analyses [539]. Fourth, a generic recommendation that may already be applicable nowadays is to deliberately turn the confirmation bias [559] automatically reinforced via AI-empowered social media [273] *against itself* (see also Chapter 2). For example, one could create social media spaces (subsuming future social VR) that reinforce *critical thinking, life-long learning and criticism* (see Chapter 2) – which could be *deliberately* fueled via artificial bots (or non-player characters in VR) steering attention towards those patterns. Even if immersive falsehood would often not be resolved quickly, an (AI-aided) social peer pressure reinforcing critical thinking and a focus on invariant good *explanations* could represent a necessarily incomplete but principled defense.

3.5 Conclusion

In this chapter we analyzed the extent to which malicious actors in AIVR could compromise truth across diverse areas from societal contexts to science. In the light of affective realism and the perceiver-dependent nature of social reality, we deconstructed the nature of the term "ground-truth" often prematurely assigned to the human experiential world. In a nutshell, we concluded that on a more strict deflationary account of truth linked to science and *not consensus*, we do *not* inhabit a post-truth era. First, humans were never equipped with a direct access to physical reality in the first place. Second, the

goal in science should in any case not consist in attempting to empirically identify and justify truth because neither positive evidence nor consensus ever establishes truth as put forth by Popper. Instead, the scientific method ideally heavily relies on *falsifiability*. In a further step, we thus analyzed whether falsifiability could be irreversibly endangered by immersive falsehood. Our analysis suggests that while the speed of falsification procedures could be considerably slowed down (which could generate serious complications in a broad range of domains including science, law enforcement, journalism and politics), it would be *a matter of degree* and not of kind. Generally, whatever level of deception and disinformation is achieved by malicious actors, it does *not* per se eradicate the scientific method and we likewise do *not* inhabit a post-falsification era. A general epistemic view on science compatible with this is to conceive of it as an endless error-corrected quest for invariant hard-to-vary theoretical explanations of reality as advocated by Deutsch [158] – a quest which can obviously *not* be terminally disrupted by slowed down experimental falsification procedures. Last but not least, we proposed to defend against and face immersive falsehood by utilizing AIVR safety tools offering a rich counterfactual experiential testbed [20, 12]. Ideally, these methods could contribute to what one could call a renewed counterfactual era of *technology-augmented critical thinking*. In short, while immersive falsehood neither terminally disrupts truth nor falsification, a technology-augmented critical thinking (and concurrently a dynamic augmentation of creativity [21] to craft novel unpredictable requisite solutions) seems indispensable in the light of various remaining severe socio-psycho-technological risks that future immersive falsehood could involve and reinforce. Future risk examples could range from AI- [94] and VR-enabled [121] crimes to false memory constructions [336, 473] over political unrest and safety-critical polarization in social media [453] (subsuming future social VR).

3.6 Epistemic Meta-Analysis

In the following, we retrospectively provide additional comments on the chapter. Firstly, we extract the key takeaways relevant for AI-related epistemic security issues. Secondly, we summarize key insights for epistemically-sensitive AI design.

3.6.1 Relevance for AI-Related Epistemic Security Strategies

In analogy to the last mainly AI-focused Chapter 2, this specifically AIVR-focused chapter simultaneously cautions society against both *over*estimating and *under*estimating present-day AI in VR contexts. On the one hand, we conclude the following: 1) the misuse of present-day AI(VR) can neither initiate nor exacerbate a "post-truth era" since the concept is inconsistent in the first place and its applicability can be refuted using

Popperian epistemology, 2) the misuse of present-day AI(VR) does also *not* confer the capacity to engender a "post-falsification era" since in theory, instead of a qualitative disruption of falsification processes, worst-case falsification complications would stay a matter of degree and not of kind. On the other hand, "immersive falsehood" in AIVR should *not* be underestimated either because a slowing down of falsification procedures yields epistemic threats affecting not only the immediate context of VR environments but also more broadly e.g. epistemic processes in science, law enforcement, journalism, politics and the collation of historical records. Given the importance of epistemological philosophy for international epistemic security that became more and more palpable in the current deepfake era, it seems vital to keep dynamically refining one's epistemic premises using ever better new explanations (and as suggested in this book the currently best format for new explanations is the special case of new *explanatory blockchains* (see Chapter 1)). Hence, epistemic security should be understood as a *process* and not the establishment of an illusionary sustainable end state. In this vein, note that strictly speaking, as elucidated by Frederick [200, 202], one must criticize statements such as the one of Popper describing that *"inconsistency and incoherence do establish falsehood"* [411] mentioned in Section 3.3. Among others, following Frederick [200], one reason for this is that a falsified theory could still be true if the accepted observations believed to have falsified it were wrong [200]. This is in line with the Duhem-Quine thesis [239] which emphasizes the context-dependency of falsification procedures as already adumbrated earlier in Section 3.3. Gradual reflections on how to build up a more robust epistemological grounding that accounts for Frederick's explanations are implicitly integrated in Chapter 5, 7 and 8.

3.6.2 Relevance for Epistemically-Sensitive AI Design

While Section 3.4 already collated a first set of epistemically-relevant AIVR design strategies as countermeasures against "immersive falsehood", the following Chapter 4 deepens this analysis by focusing on the immersive journalism use case. However, prior to that, against the backdrop of the need for epistemic refinements just mentioned in Section 3.6.1, it may be crucial to already consider its key implications for epistemically-sensitive AI design. In particular, it may now become clear that the act of selling AI products described to offer results such as "deepfake detection", "lie detection", "truthfulness" or specifically "immersive falsehood detection" would be epistemically problematic on multiple grounds. Firstly, the detection of AI-generated synthetic artefacts is embedded in ongoing adversarial cat-and-mouse games. For instance, the latter can lead to "authentic" samples being misclassified as deepfake and deepfake samples being misclassified as "authentic" through adversarial interference (see also Chapter 2). Secondly, among others due to the possibility of inaccurate observations, observed inconsistencies cannot prove falsehood. Thirdly, knowledge is steadily evolving and revised gradually by what that which is colloquially referred to as "ground truth" is an unpredictably moving target. For instance, it is

possible that something that would nowadays be perceived as scientific "fake news" may become tomorrow's agreed upon scientific knowledge and vice versa. Indeed, as emphasized by Corazza [128], there exists cases in the history of science where what was deemed to be impossible suddenly became accepted to be possible after unexpectedly being made problematic by experiment and later refuted by a new better explanation able to account for that. Fourthly, the idea to utilize AI to detect "falsehood" can worsen the stigmatization of humans being statistical outliers (see Chapter 7 for more details). In sum, we conclude that a responsible epistemically-sensitive AI design framework could implement present-day AI systems for epistemic *assistance* and cyborgnetic creativity augmentation but *not* to replace cyborgnetic epistemic judgments.

Chapter 4

Malicious Design in AIVR

This chapter is based on a slightly modified form of the publication: N. 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. 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.

4.1 Motivation

For humans to benefit from progresses in the AI field, it is essential to early start to also tackle the potential risks associated with AI development and deployment. In the light of the foregoing, AI safety and AI ethics considerations are gradually being recognized as indispensable component of AI research efforts across multiple research subfields [146, 166, 226, 326, 405, 478] at an international level [34]. Commonly, methods in AI safety and AI ethics focus on how to implement ethical and safe AI systems and how to avoid unintentional failure modes related to design-time mistakes and operational failures. However, from a cybersecurity-oriented view in AI safety [19, 84, 407, 555], it has been emphasized to *additionally* consider the existence of malicious and *unethical* actors. Such adversaries can intentionally launch malicious attacks on deployed AI systems or themselves craft AI systems with *intentional malice in design*. Concerning VR settings, recent work on security and safety for VR [404, 104, 222] and also more generally mixed reality [41, 151, 237, 512] is in line with this cybersecurity-oriented AI safety perspective and stresses the need to anticipate misuses and attacks by malicious entities. Generally, the AI risks embodied by malicious design can be understood as worst-case scenarios in AI safety given that the system is owned by the attacker allowing maximal adversarial

capabilities with minimal restrictions in white-box settings[1] [549]. Obviously, the same holds analogously for malicious design in VR.

Hence, given the promising avenues that beneficial synergies between AI and VR technologies started to bring forth [527], it seems important to proactively identify possible misuses of such synergies. In fact, the early consideration of individual use cases involving malevolent actors has been recently explicitly recommended for a security-aware development across diverse mixed reality instantiations [237]. In this chapter, we focus on malevolent design at the intersection of AI and VR (AIVR). More specifically, we zero in on intentionally performed unethical AIVR design that construes what we called "immersive falsehood" in Chapter 3. We regard immersive falsehood as a landscape of deliberately designed synthetic immersive realities for malicious purposes. For a graspable analysis and by way of illustration, we use the not yet prevalent but cogitable use case of targeted disinformation via VR news contents potentiated by Generative AI (such as e.g. via exploits using future extensions of VR deepfakes [78]). Indeed, regarding the information ecosystem in the near future, progresses in the nascent field of immersive journalism (which refers to news formats allowing participatory first-person experiences of recreated news events and situations [153]) already include the creation of VR news productions [401]. Thereby, while the convergence of AI, VR and such experiential news could provide a unique window of opportunity for innovations, it could also simultaneously exacerbate the space of possibilities for *malicious AIVR design* and disinformation.

Note that even if advanced AIVR applications might currently belong to a niche in its infancy, the analysis of this particular type of risk could be already useful *today* for AI, VR and their safety separately. The reason being that instructive insights gained from such worst-case scenario considerations might already be applicable to simpler cases. For instance, when being equipped with methods on how to defend against misleading immersive journalism experiences in VR, one might be better informed on how to tackle the consumption of manipulative disinformation videos crafted with Generative AI which are non-immersive and affect much less sensory modalities. Concurrently, a similar set of methods might be useful to sensitize VR users and raise security awareness on the procedure of potential adversaries including their manipulation techniques. Finally, the mere confrontation with instantiations of malevolent AIVR design leads involved researchers (and ultimately society) to face the relevant issue on how to possibly generate *defense methods* against vividly experienced "immersive falsehood". One possibility to address this task is the use of *design fictions* [73, 72, 112, 426] known from human computer interaction (HCI) and which have been recently also applied to near-term and long-term security issues of modern technologies including Generative AI [262]. (Co-creation design fictions can be used for technological future projections by experts in the form of e.g. nar-

[1]Simply put, in security, white-box settings refer to cases where the adversary has full knowledge about the internal implementation of the system.

ratives or construed prototypes that can be represented in text, audio or video formats but also in VR environments [406].) The next Section 4.2 collates technical and psychological information on malicious design in the context of Generative AI and immersive journalism. Then, Section 4.3 first discusses parameters relevant to a cybersecurity-oriented modelling of adversarial capabilities and goals. Subsequently, we elaborate on how to – on this basis – co-create defenses using immersive design fictions. Thereafter, we conclude in Section 4.4 and provide incentives for future work.

4.2 Malevolent Actors and Fakery in AIVR

Malevolent creativity [140] can be described as the deliberate utilization of creativity in the service of harmful goals. As in cybersecurity, malevolent creativity applied to AI may fuel incessant races between adversaries and defenders. However, as in cybersecurity, dynamic exchanges between defenders and ethical hackers and the practice of considering safety aspects, attacks and defenses can contribute to a more informed and balanced security ecosystem [407]. Early analogous efforts can be already observed in the field of adversarial machine learning where an increasing number of publications on both adversarial AI attacks and adaptive defense methods are produced [101]. Already in current AI contexts, it is technically feasible for malevolent attackers to for instance intentionally cause a variety of failures ranging from exploitation of vulnerabilities against adversarial examples over poisoning attacks and machine learning backdoors to model thefts [314]. Regarding malevolent AI design [407, 549] itself, feasible examples include among others AI-based malicious software [492], misuse of automated drones [125, 19] or autonomous vehicles [94] and *malicious design of Generative AI* [262] for disinformation, extortion and defamation.

In VR, it is in principle practicable for malevolent actors to cause psychological or physical harm [512] e.g. by displaying or overlaying offensive undesirable contents [104], enacting harassment in social virtual spaces [70], by controlling the physical movements of the user towards maliciously chosen physical locations or by deliberately inducing dizziness and confusion [104]. In extreme cases, physical harm could be caused for instance by manipulating subtle elements such as the frequency of visual stimuli threatening hereto neurologically vulnerable individuals [41]. Malevolent actors could also threaten privacy in social VR settings [383] e.g. via identity thefts of user avatars [512] linked to the unethical tracking of multiple private channels. Furthermore, an already emerging phenomenon is the unethical crafting and sharing of synthetic non-consensual VR contents [121] – which could be exacerbated with perceived virtual replica or tailored modifications of existing humans [344] if performed non-consensually. Corresponding endeavors could be fueled by future extensions of VR deepfake methods that are technically already feasible [78, 121]. Ultimately, it is conceivable that AI-aided malevolent VR design could also be utilized

e.g. for manipulative purposes at a larger scale taking the form of immersive disinformation [512]. In fact, while one advantage of the already existing VR-based immersive journalism [401] is the *"unprecedented access to the sights and sounds, and possibly feelings and emotions that accompany the news"* [153], this feature could be systematically exploited in order to deceive – especially when amplified with Generative AI. Incisively, a recent online article expounded that combining deepfakes and VR may *"damage the trust in shared information"* and could lead to *"extremely manipulated content across various channels"* [483].

4.2.1 Malicious Design of Generative AI

The currently most sophisticated instances of Generative AI that are potentially available to malevolent actors are the so-called "deepfake" techniques harnessing deep learning (DL) tools. While often associated with face-swapping methods, the range of deepfake applications transcends those contexts and comprises not only modifications of faces in image and video artifacts but also extends to speech, text and body motions as well as images in other domains. Thereby, it is important to note that deepfakes simultaneously open up a variety of beneficial and forward-looking applications (see e.g. [98] for an overview) in areas such as gaming, entertainment, health care, education or even privacy-preserving journalism. Here, we are concerned with potential misuses which if ignored, could also compromise or overshadow the unfolding of positive impacts of these technologies. Potential harmful and malicious adversarial goals to design deepfakes comprise among others disinformation, revenge, extortion, sabotage, smearing, frauds, crafting a tool for other cybercrimes, scams, impersonation, obfuscation, tempering with legal evidence and physical harm [10, 358, 516]. In the next paragraph, we introduce a set of practically relevant risk instantiations for illustrative purposes.

The following exemplary high-level processes could be instrumentalized across different domains for malicious Generative AI design: 1) *replacement*, 2) *reenactment*, 3) *image synthesis*, 4) *speech synthesis*, 5) *synthetic text generation*, 6) *adversarial perturbation* and interestingly 7) *automated disconcertion*. The most popular application for *process 1* is certainly facial replacement (aka face-swapping) in the computer vision domain. Such a DL-based facial replacement has been for instance used for a public defamation video shared across ca. 40000 individuals portraying the journalist Rana Ayyub in pornographic contexts she never partook. Concerning *process 2* which often involves a type of puppetry via facial reenactment where facial features of a driving source entity are transferred to the face of a target, they *"give attackers the ability to impersonate an identity, controlling what he or she says or does"* [516]. With increasing generative capabilities, it is easily conceivable that it could become more and more problematic for audiovisual journalistic contents. Moreover, *process 3* facilitates the generation of fake artifacts perceived as por-

traits of possibly existing individuals. It has been already utilized to generate misleading profile pictures on social media to simulate fake personas [102] and has been harnessed for disinformation [452] and even espionage attempts [451]. Another example of malicious DL-based image synthesis is the generation of deepfakes in the domain of medical imagery to add or remove diagnostic features for which a proof-of-concept has been recently implemented as applied to scans for lung cancer [359]. *Process 4* has been for instance utilized for a type of DL-based voice cloning facilitating an impersonation of the CEO of a company in the UK where an employee could be convinced to transfer a significant amount of money [486]. Very recently, *process 5* instantiated in the form of DL-based natural language generation with a fine-tuned version of the known Generative Pre-trained Transformer (GPT-2) model has been argued to be able to formulate textual messages resembling political disinformation [505].

When it comes to *process 6*, the key motivation of adding adversarial perturbations to a previously crafted material to evade deepfake detectors (a technically feasible strategy denoted "adversarial deepfakes" [370]) could be to disguise other cybercrimes or to conceal inauthentic contents related to disinformation campaigns [102]. Its future real-world instantiations could lead to severe forensic consequences [516] and could have nefarious impacts on the information ecosystem. Beyond that, it could also lead to repercussions regarding content filters related to terroristic propaganda or child abuse [125] (which is for instance conceivable if illegal authentic material is first modified via deepfakes for identity obfuscation [489] and subsequently adversarially perturbed [370] to evade deepfake detectors). Last but not least, an interim retrospective view of this short non-exhaustive enumeration of processes that can be exploited for malevolent Generative AI design reveals the need to consider the socio-psychological and forensic impacts *of their mere existence*. In fact, with *process 7* of automated disconcertion (see Chapter 2), we refer to the automatically eventuated mechanism that is brought forth by the very availability of these processes which are potentiated by the malicious Generative AI design itself. In forensics, it materializes in the form of the liar's dividend [516] seemingly taking away the general credibility of audio, visual and textual samples. At the societal and interpersonal but also intrapersonal level, it means that founded or unfounded suspicions of fakery might turn out to become harder and harder to resolve in practice. Needless to say that the rather diffuse automated disconcertion can represent a strategical advantage for malicious actors interested in forms of targeted disinformation. In fact, a recent failed military coup in the context of pre-existing political unrest in Gabon was partially grounded in the proliferation of the wrong assumption that an official presidential video represented a manipulative deepfake video [199, 235, 434].

4.2.2 Immersive Journalism, VR and Disinformation

One striking vision for the nascent field of immersive journalism (IJ) as revealed by De la Peña (who has also been called the "godmother of VR" [300]) was the explicit goal to reinstitute *"the audience's emotional involvement in current events"* [153] which seemed to exhibit a certain degree of indifference towards human suffering. It has been argued that IJ can promote empathy [449] as well as a sense of awe and wonder [402]. Moreover, a recent study found that it can foster postive attitudinal changes [89]. Further, it was initially postulated that VR news contents when *"based on 3D video rather than on 3D synthetic modeling and animation"* [153] would offer an even more realistic framing than conventional formats. This may also apply to VR content creation with modern highly detailed and realistic 3D reconstructions [162, 382, 527]. Beyond that, VR may offer *"a powerful plattform to re-create news events, taking the idea of photographic documentation of reality or acoustical recordings to an entirely new level in which the user can be virtually present at a news event and experience it as a witness or even as a participant"* [401]. Interestingly, VR news experiences have been associated with higher telepresence and even elevated news credibility [281] when compared with standard news consumption forms without VR exposure. IJ experienced with VR headsets could allow unique experiences of immersive 3D "spatial journalism" [402] via *"the introduction of user-directed spatial dynamics, adding a new level of presence"* [93]. Despite these promising avenues and the fact that there exist multiple types of IJ including AR frameworks, 360-degree reports [238] and drone-based immersive news [402], IJ is still in its infancy and the most widespread pieces correspond to either 360-degree video productions or mobile VR settings [88, 401] which is also linked to the fact that VR content creation is still relatively complex and expensive nowadays [343].

Nevertheless, multiple early IJ formats in VR have been developped in the last two decades. The first VR news story of the New York Times (NYT) (albeit only as 360-degree film downloadable from a NYT app which some would strictly speaking not label as VR content [500]) termed "The Displaced" [401] was focused on three children from different nations displaced by war and allowed a visual exploration of the effects of the devastation. Furthermore, "Project Syria" facilitated an immersive VR experience featuring a bomb explosion in Aleppo and a refugee camp [401] that could be viewed with Oculus Rift or HTC Vive while "Assent" was devised as a VR documentary that could be viewed with Oculus Rift depicting the witnessing of military executions in Chile from the perspective of the maker's father. Another IJ piece in VR that was made available to the public was crafted to raise awareness concerning the detention conditions at the Guantánamo Bay prison and was based on a re-construction of this prison for Second Life and later for Unity3D. In these examples of VR journalism, a unique grasp of the situation becomes possible by *"transferring people's sensation of place to a space where a credible action is taking place that they perceive as really happening, and where, most*

importantly, it is their very body involved in this action" [153]. In a nutshell, according to De la Peña, it is this combination of presence, the plausibility of the experience and the embodied active sampling of the environment that facilitates this *"profoundly different way to experience the news, and therefore ultimately to understand it in a way that is otherwise impossible, without really being there"* [153].

However, this set of attributes of IJ in VR make it at the same time lucrative for malicious actors. It is easily conceivable that such unique immersive experiences can also create presence, immersion, empathy and a sense of credibility in the context of fakery advocated by manipulative entities [277]. Different IJ formats could accordingly be misused for propaganda and disinformation. For clarity, instead of using the broader term of "fake news" (which partially overlaps with disinformation and misinformation [324]) to refer to misleading information and news contents, we utilize the narrow term "disinformation" in the sense recommended by the UK government. It defines *"disinformation as the deliberate creation and sharing of false and/or manipulated information that is intended to deceive and mislead audiences, either for the purposes of causing harm, or for political, personal or financial gain"* [122]. Regarding disinformation in IJ, in a 2017 interview with Quartz, De la Peña stated that *"VR will be used for propaganda. It will be used badly for journalism. [...] But that's always going to be about, who's the maker? And it's not about the medium"* [418]. For this reason, the main concern addressed here is malevolent creativity exhibited by malicious makers. (Naturally, other concerns may stem from unintentional human errors.) As it has long been the case in cybersecurity and now also in deepfakes, there is certainly an attacker-defender arms race when it comes to disinformation attempts. Future IJ and also VR itself could arguably follow this type of trend [237].

4.2.3 Manipulated VR News and False Memory Construction

In short, with VR technologies becoming cheaper and more widespread, "immersive falsehood" fabricated by malicious actors could emerge in IJ settings. Sanchez described related possible dystopian scenarios *"where users are immersed in a world of fake news"* [449] while Uskali and Ikonen [511] stressed that IJ experts should be aware of *"[...] advanced and sophisticated manipulation and disinformation operations [...]"*. Beyond that, Uskali et al. specify that *"our brain believes so strongly in what it sees in VR that we might not be able to distinguish fake news from real news"* [510]. In our view, one very specific concern for the future of IJ is the targeted and tailored elicitation of false memory constructions via experiential VR news contents. On the one hand, when compared to traditional desktop displays, it is known from recent findings that immersive VR with head-mounted displays affords *more memorable* experiences by combining *"visually immersive spatial representations of data with our vestibular and proprioceptive senses"* [311]. On the other

hand, this concise feature of long-lasting effects via the spatially-centered experiences in VR journalism [93] could open up a novel attack surface for malevolent actors interested in disinformation operations. More generally, Liv and Greenbaum [336] postulate that *"creating false memories to promote the uptake of fake news, both on the individual and mass scale can be enabled through multiple different means, including narrative, video, photos, and virtual reality"*. In this line, a study of Frenda et al. emphasizes that fake memories can be specifically brought forth for manipulative political gain – with successful uptakes especially if the contents are coherent with pre-existing preferences [203]. Another study found that elementary-aged children are susceptible to false memory formation in VR [464] and concluded more broadly that *"third parties may be able to elicit false memories without the consent or mental effort of an individual"*. Already the exposure to a small set of misleading photographs and a narrative led to false memory construction across half of the adult participants in a 2018 study in the period preceeding the Ireland abortion referendum [367]. Overall, it is easily conceivable that hyperrealistic IJ pieces experienced with VR headsets may exacerbate such psychological effects[473].

Against the backdrop of the foregoing analysis speaking to the creation of *durable false memories* for disinformation purposes, the following exemplary set of 3 processes could facilitate this endeavor: 1) *persuasive spatial dynamics engineering*, 2) *memory-centered sensory stimulation*, 3) *information gathering*. *Process 1* refers to any set of systematically selected processes whose outcome yields increased *spatial* awareness, perception and orientation in VR settings (such as e.g. implementing 3D minimaps [305]). *Process 2* consists in selecting any specific sensory stimulation that increases memory consolidation. For instance, it is easily conceivable that future adversaries could especially profit from the already implemented [368, 373, 354, 502, 421, 528] but not yet available-for-sale *olfactory* displays for VR. The reason being that neuroanatomically speaking, olfactory pathways are unique [488] and olfactory memories differ from other memory forms by being particularly apt to evoke affectively-loaden memories and having a strong propensity to influence memory acquisition while being at the same time *"highly resistant to forgetting"* and *"highly resistant to retroactive interference"* [435]. (Olfactory displays can for instance be attached to VR headsets [368, 421] or utilized as on-face [528] or handheld [373] wearables.) Finally, *process 3* could include various techniques such as e.g. social engineering or open source intelligence gathering [42] (retrieving publicly available data on a target) to *identify pre-existing preferences and beliefs* of the victims to be able to match VR contents for a successful uptake of disinformation.

4.3 Cybersecurity-oriented Immersive Defenses

In the last Section 4.2, we reflected upon the space of affordances available to potential malevolent actors in AI and VR respectively. We illustrated this concept utilizing the use case of disinformation in immersive journalism contexts. In this section, we discuss a cybersecurity-oriented methodology to generate defense methods against adversaries operating in AIVR, at the intersection of AI and VR. In this vein, in cybersecurity and also in recent work on security for machine learning, it is indispensable to perform a so-called *threat modelling* [101], a clear specification of assumed goals, capabilities and knowledge exhibited by the adversary. For this reason, prior to elaborating on how to generate generic immersive AIVR defense measures, we first provide a threat model for our malevolent AIVR design use case for illustration.

4.3.1 Threat Modelling for Malevolent AIVR Design Use Case

- ***Adversarial goals:*** Given the choice of our use case, the goal of the assumed adversary is a specific form of targeted disinformation by combining AI with VR tools in IJ settings. We consider that the adversary has the specific goal to manipulate the opinions, attitudes and views of selected IJ victims in a well-defined manner according to a self-defined scheme. More precisely, the goal could be to modify a source set of conceptions S to a target set of conceptions T in a certain context whereby these sets could differ in content and in confidence assigned to each element. By such a modification, the adversary intentionally aims at deceiving and misleading based on political, personal or financial motives or/and as an end in itself to cause harm. Overall, the adversarial goal would correspond to a *microtargeted disinformation* in IJ.

- ***Adversarial knowledge:*** We assume that all Generative AI components utilized are available in *white-box* settings. The same holds for the VR content creation for the IJ experiences that is fully transparent to the adversary. Moreover, the adversary is able to gain publicly available information on the victims and can attempt to gain more personal data via social engineering. One can conceive of malicious Generative AI (and by extension malicious deepfakes and VR deepfakes) as a type of adversarial examples on humans – as exposure to a specifically arranged sensorium with the goal to fool human entities (at the level of their preferences, beliefs and perceptions). Hence, in the case of humans for which information gathering succeeded in supplying crucial personal knowledge, it may be described as *grey-box* setting (a nuance between black-box and white-box adversarial knowledge levels).

- ***Adversarial capabilities:*** Regarding the Generative AI parts, the adversary can

at least instrumentalize the set of 7 processes introduced in the last section which consisted of replacement, reenactment, image synthesis, speech synthesis, synthetic text generation, adversarial perturbation and automated disconcertion. In the VR content creation, the subtasks relevant to the disinformation goal are under the control of the adversary. For instance, we assume no constraints on the design and combination of the multimodal material for content (e.g. images, videos, audio samples,...). The adversary has no constraints on performing the 3 mentioned processes for VR content creation: persuasive spatial dynamics engineering, memory-centered sensory stimulation and information gathering. Thus, in total, the adversary can leverage 10 different processes to achieve microtargeted disinformation. However, it is obvious that in practice the set of capabilities could be wider and is solely constrained by malevolent creativity which is why defenses should be understood as incremental techniques and not as conclusive solutions.

4.3.2 Immersive Design Fictions for AIVR Safety

Design fiction (which we abbreviate with DF in the following) enables *"HCI and design researchers to co-create, explore and speculate the future"* [6]. Very recently, Houde et al.[262] successfully applied co-creation DF to the specific context of near-term AI safety related to (mis)use cases of Generative AI. On this basis, we regard DF as a well suited methodology for defenses against near-future AIVR safety risks as illustrated in this chapter. For clarity and to facilitate a systematic procedure, we suggest to ground future AIVR DF endeavors for defenses in threat models. Moreover, the law of requisite variety in cybernetics suggests that *"only variety can destroy variety"* [32]. Applied to our use case, this signifies that in order to identify requisite knowledge for defenses against the described threat model of an adversary operating at physical, virtual and importantly immersive levels, one may profit from an immersive perspective. In our view, this need for an immersive stance for the meaningful generation of solutions applies generally to any malicious AIVR design use case linked to "immersive falsehood". Interestingly, it has been proposed to utilize VR as a powerful platform for DF given the *"higher level of immersion and sense of embodiment"* [6]. In a nutshell, *AIVR safety can profit from AIVR* (next to multiple other areas such as e.g. cybersecurity, social psychology, affective science, law or journalism) and vice versa.

In the light of our threat model, it becomes clear that DFs for such malicious AIVR design use cases need to consider a *socio-psycho-technological* threat landscape with immersive, digital and physical elements and profound cognitive-affective implications. Given the complexity, a meaningful approach requires *transdisciplinary* dynamics. Importantly, the DFs need to not only address proactive defenses, but also *reactive* mechanisms [19]. In fact, proactive defenses could aim at hindering malevolent actors in AIVR to be able

to disseminate their VR contents in the first place. Such measures could for instance include prevention mechanisms *preceding* content deployment and could be developed based on tools analogous to deepfake detection AI. However, given the fallibility of human knowledge, the unreliability of AI detection systems and the unpredictability of human malicious creativity, one needs to be aware of the need for reactive defense measures i.e. in the example of our use case *after* users were exposed to the manipulated VR news contents.

Notably, we do *not* consider DF as a tool to *predict* the future. Given the unpredictability of future knowledge creation, future extrapolations are limited by the state of available present knowledge and reactive measures to unknown unknowns will be needed. DF cannot foresee the consequences of not-yet created knowledge. However, DF allows the generation of plausible counterfactual paths that *could* become crucial. Organizationally, we assume a preparation phase *preceding* the DF in which an *immersive prototype* is crafted (more details below). A simple prototype could e.g. be an immersive multimodal storytelling narrative with audiovisual (see e.g. recent MIT deepfake storytelling project [361]), olfactory or tactile material. For the future, we ideally recommend a VR prototype [484]. Overall, we consider 3 disjunct groups: the makers of the immersive prototype, a set of designers with expertise in AIVR and a multidisciplinary set of participants with knowledge in a variety of technological areas overlapping with AI and VR or not. The following order for the immersive DF is non-binding and has a merely illustrative function:

1. ***Designer co-creation session:*** A group of AI and VR designers craft a *threat model 1* and a *threat model 2*. The former refers to a use case of a malicious AIVR design that would already be technically feasible nowadays and the latter to a use case that they consider feasible in 5 years given their current knowledge.

2. ***Participant introduction to AIVR:*** The AI and VR designers provide a high-level introduction to the multidisciplinary audience. It provides an overview on the state-of-the-art of technical possibilities at the intersection of AI and VR.

3. ***Designer narrative:*** The designers present *threat model 1* to the audience.

4. ***Participant co-creation session:*** Instructed by this example, the participants generate a new *threat model 3* based on what they assume might be technically feasible in 5 years given their current knowledge.

5. ***Participant narrative:*** The participants present *threat model 3* to the designers.

6. ***Narrative comparison:*** The designers present *threat model 2* and participants compare it to *threat model 3*.

7. ***Immersive session:*** Designers and participants undergo a short experience of the immersive prototype. The prototype experientially conveys a *threat model 0* (prefabricated by the makers of the prototype). In our use case example, it could consist of a short *blind* immersive experience with 2 pieces: an IJ piece (ideally in VR) featuring an *unknown but real event* and another one featuring disinformation inspired by the threat model in Subsection 4.3.1. Before and during exposure, users are not informed on which piece is real and which manipulative. Clarification is provided at the end.

8. ***Common defense co-creation session:*** Designers, participants and makers co-create proactive *and reactive* defenses against threat models *0* to *3*. They also discuss possible adaptive attacks (when defenses are known).

4.4 Conclusion

Recent research related to the safety and security of AI and VR respectively emphasizes the need to complement classical efforts to design *ethical* and safe systems with the anticipation of intentional exploits by *unethical* and malicious actors. In this vein, we performed a proactive cybersecurity-oriented analysis of malicious design in AIVR i.e. at the intersection of AI and VR. Even though the field is in its infancy, it is essential to build more robust dynamics *from the onset on* [512] and not in hindsight. By way of illustration, we applied our analysis to the use case of immersive journalism where malevolent actors could specifically harness Generative AI and VR settings for purposes of (microtargeted) disinformation creating "immersive falsehood" – with socio-psycho-technological implications that may require proactive and reactive *immersive defenses*.

For the purpose of generating such defense measures, we introduced a cybersecurity-oriented approach to immersive co-creation design fictions (ideally in VR). In a nutshell, *AIVR safety can benefit from immersive AIVR co-creations*. Thereby, while such co-creations may not represent a panacea to counter malicious design, it seems recommendable to incrementally employ and update them on-demand for conceivable AIVR safety use cases. Beyond that, it can be postulated that immersive design fictions inspired by security practices represent a possible way to utilize VR as rich counterfactual experiential testbed [12, 20] – however now extended to counterfactuals comprising co-existing *unethical* actors.

In a recent futures exercise, AI-generated fake content was ranked among the highest-rated potential applications for AI-enabled crime [94]. Moreover, Generative AI such as deepfakes could be used for the malicious creation of false memories [336]. Such considerations paired with the aptness of VR to facilitate durable memories represent AIVR

synergies that could be exploited by malicious actors. The possible psychological implications of false memories induced in the context of such exploits could be studied in future work. Thereby, a promising avenue for future prevention and remedies could perhaps also include immersive cognitive-affective debiasing measures harnessing AIVR itself.

4.5 Epistemic Meta-Analysis

As conducted in previous chapters, we retrospectively add comments to contents of the chapter. Firstly, we extract the key takeaways relevant for AI-related epistemic security issues. Secondly, we summarize key insights for epistemically-sensitive AI design.

4.5.1 Relevance for AI-Related Epistemic Security Strategies

As can be extracted from this chapter, when it comes to epistemic security considerations pertaining to the information ecosystem in the deepfake era, it is crucial to *additionally* cover social VR – especially in light of its growing popularity among users. The chapter focused on the exemplary epistemic threat that maliciously crafted deepfake-augmented immersive journalism settings could pose to society. However, we conclude that much more generally, one should not *under*estimate the variety of novel possible synergies between AI and VR that could be deliberately exploited by malicious actors for purposes of epistemic distortion. Thereby, the construction of what has been referred to as "false memories" in psychology which could be exacerbated in VR exemplifies the insufficiency of an empiricist approach to epistemology in the deepfake era. In light of the epistemic stance conveyed in this book which foregrounds the phenomenon of new explanatory blockchains (EBs) (see also Chapter 1), it is apparent that because the creation of any non-EB-like information can be forged, one can indeed *not* rely solely on one's non-EB-like memories to stay epistemically secure in the long-term. The importance of EB-based strategies for AI- and also VR-related epistemic security is discussed in Chapter 7.

4.5.2 Relevance for Epistemically-Sensitive AI Design

Interestingly, an epistemically-sensitive AI(VR) design could craft immersive design fictions for the virtual exploration of EB-based strategies (see also Chapter 7). One could state that a remarkable property of the best scientific, technological but also philosophical EBs is that by the comparatively strict nature of their generation procedure, (think e.g. of detailed texts describing new physical theories or new patents all of which could at least be easily transformed into a suitable EB format) belong to the most rigid invariant epistemic artefacts. Changing minute semantic details engenders inconsistencies that

could undermine the validity of the entire chain of explanations – which would not correspond to an EB anymore. To shed a new comparative light on the term "hard-to-vary" utilized by Deutsch [158], one can state that EBs are fundamentally hard*er*-to-vary than any *non*-EB-like information[2]. In that sense, the epistemic procedure according to which explanation blocks are connected within one EB makes the latter semantically immutable. Moreover, a meta-chain of successively discovered new EBs that improve upon each other may inherit that immutability[3]. In a nutshell, one could utilize this peculiarity as a feature for an epistemically-sensitive AI design facilitating a technology-augmented critical thinking. For example, one application could be to let humans experience the contrast between the contents of a new EB on a problem x versus new non-EB-like explanations on the same problem x with the required new non-EB-like material being *deepfake text* generated by a language AI. (Obviously, it is also possible for a hereto willing human to manually generate such new non-EB-like material.) More details in this regard are provided in Chapter 6 and 7. Prior to that, the next Chapter 5 first focuses on a new epistemic threat emerging in the deepfake era: scientific and empirical adversarial attacks, a new form of malicious AI-aided epistemic distortion of which "deepfake science attacks" represent a subset.

[2]This can be interpreted as one major reason for the present-day lucrativity of (cyber) intellectual property theft by malicious actors.

[3]Nowadays, one could argue that in epistemically-relevant areas such as science, anthropic peer review in a domain would ideally already *implicitly* implement a human expert peer network where each individual somehow instantiates (via own private expert memories) such a meta-chain of consecutive EBs given that domain in which that individual is an expert. Perhaps, to ease an AI-based retrieval of known EBs that augments humans in domains where they happen *not* to be experts, one could then implement public human-peer-review-based (i.e. more generally Type-*II-only*-peer-review-based instead of automatable such as e.g. conventional proof-of-work schemes) blockchain solutions that *explicitly* document the implicitly evolving meta-chain instantiated by human experts for each domain. Obviously, the latter could however allow no statement about novel yet unknown EBs. Yet, it could help strengthening an awareness of what humans perceive as "new" in the first place. For instance, for reasons of epistemic security, any knowledge that can be derived merely by deduction from old EBs (e.g. including with the help of present-day AI) should *not* be considered as new EB in science (see also Chapter 6 for more details). Indeed, one may even need to proactively explicitly augment currently accepted known EBs with Type-II-only-validated but Type-I-AI-generatable new *non*-EB-like information that those old EBs already seem to entail. The implementation of corresponding Type I AI and the expert validation process could yield an efficient cyborgnetic creativity augmentation method.

Chapter 5

Scientific and Empirical Adversarial (SEA) AI Attacks

This chapter is based on a slightly modified form of the publication: N. Aliman and L. Kester. Epistemic defenses against scientific and empirical adversarial AI attacks . In *Proceedings of the Workshop on Artificial Intelligence Safety 2021 co-located with the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI 2021), Virtual, August, 2021.*, 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.

5.1 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 [29, 80, 92, 186, 327], AI ethics and AI governance issues [196, 275, 386, 424] gained momentum at an international level. In addition, cybersecurity-oriented frameworks in AI safety [84, 408] 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 [565], AI simultaneously provides new affordances for attackers [33] from AI-aided social engineering [465] to AI-concealed malware [298]. Next to the capacity of AI to extend classical cyberattacks in scope, speed and scale [279], 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 [113, 279, 505] which is especially relevant to information warfare [242]. Recently, the weaponization of Generative AI for information operations has been described as *"a sincere threat to democracies"* [243]. 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" [571] 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 [346] 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) [425]. The authors utilized a fine-tuned version of the GPT-2 language model [419] 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 [524] and with the pre-trained GPT-3 language model [82] that *"may very well pass peer review"* [155] but also AI-generated fake reviews (with a fine-tuned version of GPT-2) apt to mislead experienced researchers in a small user study [490]. Future malicious actors could deliberately breed a large-scale agenda in the spirit of *"fake science news"* [252] and AI-generated papers that would widely exceed in quality (later withdrawn) computer-generated research papers [513] 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 5.3.1, 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 treated in Section 5.2.

5.2 Theoretical Generic Epistemic Defenses

As reflected in the law of requisite variety (LRV) known from cybernetics, *"only variety can destroy variety"* [32]. 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 [519] to AI safety [14] over AI ethics [31]. 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 [524] has been often thematized and even lately applied to "fake news" in the healthcare domain [46]. 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"* [77]. Another analysis presented a set of defense measures against the spread of deepfakes [113] 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"* [252], 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"* [347]. Following the author, scientists could equip citizens with sense-making tools without which *"emotions and beliefs that pander to false certainties become more credible"* [347].

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"* [82]) 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"* [194]. 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* [176]. Thereby, it is assumed that *"deepfakes reduce the amount of information that videos carry to viewers"* [176] 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 5.3 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 [87] – 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 (see Chapter 3). 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"* [412]. Thirdly, the epistemic aim in science can neither be truth directly [200] 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 [411]. Science is quintessentially *explanatory* i.e. it is based on explanations [158] and *not* merely on data. While the epistemic aim cannot be certainty or justification (and *not* even "truer explanations" [200][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 [200]. 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

[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 [200].

an explanation 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 [160]. 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"* [160].

Given 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"* [200]. 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"* [200] if the accepted observations believed to have falsified it were wrong. Thirdly, considering the 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 [445]. However, on epistemic grounds elaborated in-depth previously [14] which is amenable to experimental problematization [16], we assume that the task to consciously *create and understand* novel yet unknown *explanatory* knowledge [158] – which humans are capable of performing *if willing to* – cannot be learned by AI systems *by mere imitation*. In this book, we postulate more generally that it is impossible for Type I AI to create new explanatory blockchains (see Chapter 1 and6).

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*" [252], 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* [49]. 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***

5.3 Practical Use of Theoretical Defenses

In the following Subsection 5.3.1, we briefly perform an exemplary threat modelling for the two specific use cases introduced in Section 5.1. 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 (see Chapter 2). In Subsection 5.3.2, we then derive corresponding tailor-made defenses from the generic characteristics that have been carved out in the last Section 5.2 while thematizing notable caveats.

5.3.1 Threat Modelling for Use Cases

Use Case Security Engineering

- ***Adversarial goals:*** As briefly mentioned in Section 5.1, 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 [425]. 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"* [425]. 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 [63, 392] 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. 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 [346] 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 [415, 242]. 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 [273]. Naturally, when logically applying the epistemic stance introduced in the last Section 5.2, 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* [101], 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* (see Chapter 2) 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.

5.3.2 Practical Defenses and Caveats

As is also the case with other advanced not yet prevalent but technically already feasible AI-aided information operations [242] and cyberattacks targeting AIs [243], consequences could have ranged from severe financial losses to threats to human lives. Multiple socio-psycho-technological solutions including the ones reviewed in Section 5.1 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 5.2 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* [425]. 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* [248] as signals to calculate when it will epistemically seek clarification from a

[3]Such a system could instantiate *technical* self-awareness [14] (e.g. via active inference [476]).

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 [297] *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"* [124] 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 new but explanatorily insufficient contents. 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 [356] tailored to these attacks and complemented by *adversarial* retrospective counterfactual risk analyses (see Chapter 2) 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. 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* [308], one could perform red teaming, penetration tests and (adaptive) attacks employing AI-generated "fake data and experiments", "fake papers" and "fake reviews" [490]. Candidates for a blue team are e.g. reviewers and editors. Concurrently, urgent AI-related plagiarism issues arise [155].

5.4 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.

5.5 Epistemic Meta-Analysis

Again, we briefly retrospectively contextualize the chapter within the body of the book.

5.5.1 Relevance for AI-Related Epistemic Security Strategies

In this chapter, we started to develop epistemic defense strategies against SEA AI attacks affecting two particular examples of application areas: security engineering and scientific writing. Thereby, the three key generic features harnessed to craft those defenses (see Section 5.2) represent a starting point for the following Chapter 6. There, we summarize the premises of a *cyborgnetic epistemology* explicitly linked to the epistemic artefact of novel explanatory blockchains (EBs) – which were already mentioned in previous chapters but not yet explicitly integrated in the narrative of the current chapter. In brief, Chapter 6 explains how cyborgnetic epistemology can be used as framework to craft more robust epistemic security strategies against SEA AI attacks especially affecting *science*.

5.5.2 Relevance for Epistemically-Sensitive AI Design

As hinted in Section 5.4, one can design present-day AI to enhance threat modelling (for more details, see Chapter 7). Chapter 6 unifies cyborgnetic epistemology and *cyborgnetic creativity augmentation* to improve epistemic security in science and education.

Chapter 6

Generic Cyborgnetic Defenses Against SEA AI Attacks In Science

Chapter 6.2 is partially based on extracts from the publication: N.-M. Aliman and L. Kester. "Immoral Programming: What can be done if malicious actors use language AI to launch 'deepfake science attacks'?". *Wageningen Academic Publishers*, (2022): 179-200, 2022. 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.

6.1 Cyborgnetic Epistemology

6.1.1 Motivation

Given that human malicious actors that would decide to craft SEA AI attacks that could affect science are unpredictable explanatory knowledge creators, one is not able to remotely control the adversarial disturbances they could cause ahead of time. Hence, a risk averse solution cannot be the only option to defend against such SEA AI attacks that would affect the scientific enterprise. Consequently, instead of shielding oneself from deepfake text, which could in the long term even necessitate a retreat from society, another strategy could consist of proactively building up resilience by actively seeking *more* exposure to deepfake texts (albeit at a self-defined pace in a self-selected setting). However, for such a solution to be workable, a robust epistemology is required that does not entail justification-related epistemic threats [176] according to which the deepfake-permeated world gradually loses relational meaning via a quantitative decrease in information content. Despite epistemic dizziness, which has always existed for humans even before the

advent of deepfakes (see also Chapter 3), explanatory-anchored science cannot be terminally disrupted by additional deceptive deepfake data. Instead, when faced with deceptive material such as that produced in SEA AI attacks, one can focus on ever better explanations of the world and criticise the perceived contents on a comparative basis without having to consciously update any latent probabilistic credence. Metaphorically speaking, better explanations are our only – though ephemeral – stones on our trajectory through the deep sea of doubt. Experimental falsification shapes this trajectory but does not determine it. Explanatory-anchored science makes pragmatic progress via incremental small steps from stone to stone, which is why the epistemic aim is of a relational and comparative nature. One does not epistemically fall deeper than on one's own stones (compared to the threatening void in which a justification-based epistemology could potentially fall in times of deepfake and fake news [176]). The aim is not to find isolated good explanations, but to identify ever *better* new explanations [200] according to criteria agreed upon with others. In this book, we postulate that in the deepfake era, we must raise the bar for better explanations (see also Chapter 1). More specifically, we state that candidate better explanations must *at least* fulfil the format of new explanatory blockchains (EBs). In this vein, in the next Section 6.1.2, we formulate a new cyborgnetic epistemology that is sensitive to the epistemic security challenges of the deepfake era.

6.1.2 Epistemic-Security-Aware Epistemological Grounding

- The epistemic modus operandi is *EB-anchored*, *trust-disentangled* and *adversarial*.
- The epistemic aim is to create ever better new EBs (see example in Figure 6.1).
- A "trust-disentangled" modus operandi signifies that the epistemic modus operandi is grounded in agreed upon criteria for new ever *better* and *not* e.g. "more trustworthy" EBs. This means that it is orthogonal to any trust relation between involved entities. A better EB must be formulated such that metaphorically speaking it appears to defend itself against adversarial candidate EBs.
- An "adversarial" modus operandi signifies a conscious fallibilism at all levels. Firstly, experiments never conclusively falsify a currently accepted (i.e. instated) EB, they make that old EB problematic. We call it *experimental problematization*. Secondly, for a *(provisional) refutation* of an EB, one needs at least a new better EB. All refutations are provisional by design and can be repealed retrospectively (see also Chapter 5). Thirdly, consistent with Frederick [202], it is both rational to act *in accord* with currently instated EBs *and* to experimentally act *against* those. One reason for the latter is that experiments could at any time unexpectedly make even the best tested old EBs problematic. Moreover, to act against old instated EBs could stimulate one's creativity in crafting novel better EBs that refute those.

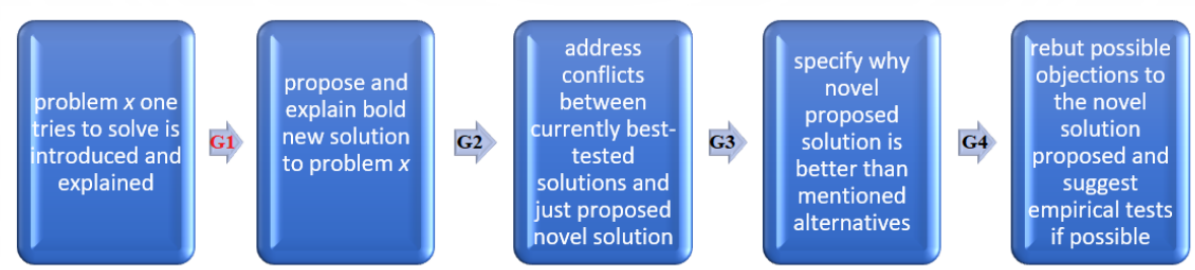

Figure 6.1: Exemplary epistemic total order for the generation of new EBs (the instructions are loosely inspired by an essay of Frederick [201]). Each glue operation x is indicated via a label G_x. EBs are a special form of explanatory information (EI) obtained by interweaving EI blocks via the step-by-step application of rational procedures sampled from a robust explanation-anchored, adversarial and trust-disentangled epistemology. In science, the specification of (direct or indirect) empirical tests in G_4 is the default condition.

- The inherently comparative criteria for "better" EBs are *updatable* and determined by agreement. Current criteria accepted in science encompass e.g. a preference for explanations that are simpler, provide more novel problematizable predictions, are more innovative, more aesthetically appealing than rival ones (see also Chapter 5).

- The inherently comparative criteria for "new" EBs must be udaptable and determined by agreement. As displayed in Figure 6.1, for a candiate EB to be accepted, the novelty criterium must inherenly be fulfilled. Indeed, the glue operation G_1 is formulated as follows: "propose and explain bold *new* solution to problem x". (Note that we must presuppose that problem x is a genuine problem [200] in the first place; not all questions are epistemically-relevant.) In the deepfake era, novelty must be adapted to exclude forgery by even the most advanced Type I AI.

- In a new EB (be it in the science domain or in philosophy), one must specify a *new* solution to a genuine problem x which must fullfil the following two necessary conditions: 1) the solution can be represented as a set of explanations S_E *and* 2) that set S_E contains *at least one* explanation $e_{Mysterious}$ that could *not* have been *reliably* generated with arbitrary high accuracy via an automatable (i.e. Type-*I-only*) pipeline given existing knowledge. The latter implies the following: given publicly available knowledge S_{OldEBs} and the genuine problem x as inputs, it would be impossible for a Type I AI to reliably generate $e_{Mysterious}$ as output.

- In the science domain, a new EB must additionally fullfil the following third and fourth necessary conditions. The third necessary condition is that the set S_E entails *at least one* new prediction $p_{Mysterious}$ for which it holds that: a) it is in theory amenable to experimental problematization but has not yet been made problematic by experiment in practice *and* b) it could *not* have been *reliably* generated with arbitrary high accuracy via an automatable (i.e. Type-*I-only*) pipeline given the set

S_{OldEBs} of currently instated old EBs. This implies the following: given publicly available knowledge S_{OldEBs}, it would be impossible for a Type I AI to reliably generate $p_{Mysterious}$. Finally, consistent with Frederick [200], the fourth necessary condition is that this prediction $p_{Mysterious}$ could *not* have been deduced *without* combining *all* elements from the set of explanations S_E.

- In the philosophy domain, one can accommodate for such lines of thought by specifying that a new EB must also fulfill the necessary condition that it is currently *not* possible to identify a subset S_{SubE} such that $S_{SubE} \subsetneq S_E$ with S_{SubE} being already *sufficient* to solve the problem x. The latter avoids superfluous statements.

- While it holds that 1) Type I AIs can in theory forge the creation of any new non-EB-like information including texts widely perceived by humans as "novel explanations", it holds that 2) due to a gap of understanding, it is impossible for all Type I entities to reliably create new yet unknown EBs respecting an epistemic total order stemming from a rigorous epistemology as e.g. exemplified in Figure 6.1.

- An experimental problematization of cyborgnetic epistemology would e.g. be a shortcut via a Type I AI able to reliably create new EBs with arbitrary high accuracy.

- A (provisional) refutation of cyborgnetic epistemology would be a better new theory that explains why such a Type-I-shortcut is possible.

6.2 Cyborgnetic Creativity Augmentation

6.2.1 Motivation

In light of the last Section 6.1.2, it becomes apparent that next to a more robust cyborgnetic epistemology, a proactive self-paced exposure to *adversarial* patterns (challenging the currently instated EBs) and other creativity-augmenting epistemic artefacts may be helpful in building resilience against SEA AI attacks. In the next Section 6.2.2, we address the question of how to implement cyborgnetic creativity augmentation in a pragmatic framework, compiling insights from creativity research in the fields of psychology [501] and cognitive neuroscience [164, 165]. The ambiguously designated *artificial creativity augmentation* research direction [21] has recently been put forth for the purpose of implementing generic defenses against societal level harm. It unifies two complementary and moreover interwoven research directions: (1) the artificial augmentation of human creativity; and (2) the augmentation of artificial creativity. Noticeably, artificial creativity augmentation represents one possible instantiation of cyborgnetic creativity augmentation. To extend our generic defenses against SEA AI attacks with the use of creativity-augmenting language models (LMs), the twofold task can be exemplarily reformulated as

follows: (1) augmenting human creativity using LMs; and (2) augmenting artificial creativity in LMs via humans. The former and the latter are intertwined since the subtask: (1) can reinforce the subtask (2) and vice versa.

6.2.2 Use Case Epistemically-Sensitive Language AI Design

Theoretical Solutions

In the spirit of recent work by Mick Ashby [31] at the intersection of cybernetics and AI ethics, one could state that in this case, humans and LMs reciprocally become a sort of intra-cyborgnetic ethical regulator of each other with the feedback loop instantiated for the purpose of counteracting unethical practices of deliberate disinformation in the (applied) science domain. Hence, cyborgnetic creativity augmentation proposed initially for security reasons against SEA AI attacks is also a form of augmenting intra-cyborgnetic ethical regulation. This in turn suddenly unifies moral programming and security research to counter immoral programming. It seems that security and ethics converge whilst counteracting SEA AI attacks. In the following, we now specifically map out two clusters of generic cyborgnetic creativity augmentation strategies. The first cluster concerns generic strategies to augment anthropic creativity using LMs. The second cluster pertains to generic strategies for the augmentation of artificial creativity within LMs. To this end, we select suitable starting points based on the ten provisional available artificial creativity augmentation indicators [21] which were grounded in explanations from psychology [501] and cognitive neuroscience [164]. Here we focus on LM-applicable options. Firstly, in order to augment human creativity using LMs, suitable generic strategies could be to design these AIs with the following enhancing subgoals: (1) increase human criticism abilities [501]; (2) stimulate human divergent thinking [190]; (3) alter the nature of self-experience at waking time [39, 253]; (4) extend the nocturnal unconscious and/or dream-related creative generation and active forgetting processes [99]; (5) encourage frequent human engagement [471]; (6) provide human sensory extension [12, 21]. Secondly, concerning the human-performed augmentation of artificial creativity within LMs, we add the following generic strategy; (7) immersion in the human affective niche via a mathematical approach and via active sampling.

Practical Use of Theoretical Solutions

- **Epistemic Context Data and Experiments:** Here, we describe how LMs could be used for an epistemically-sensitive creativity augmentation in science, education and other areas where empirical research is prominent. Creativity can be described

as a tripartite evolutionary affective construct with three modes [165, 163]: the deliberate mode (when consciously engaging in creative deliberations), the spontaneous mode (an unconscious process whose creative end result presents itself spontaneously to consciousness), and the flow mode (when creativity is enacted directly in emulations of the motor system). We focus on the two first modes in what follows. LMs could be utilized frequently to stimulate divergent thinking in the deliberate mode by first letting the scientist prompt the LM on providing a solution to a given practical problem. Since LMs are not able to create new EBs, the scientist could then criticise the generated output and re-prompt the LM, derive inspiration from it, or utilize it to question own prior assumptions. Generally, to improve the required critical reasoning abilities, a novel systematic *LM-based adversarial educational tool* could be made publicly available. In order to evade the epistemic threats of experimental justificationism that is compromised in the deepfake era, science could more widely opt for the already registered reports [380, 533] in which experimental research is assessed at an earlier stage based on the *explanatory* quality of the research proposal itself and not on the lucrativity of later documented experimental results. Building on that, LMs could then be utilized for life-long learning and for students in engineering and science to train the formulation of better EB-anchored empirical research proposals which could also include the writing of registered reports. For instance, given a current paragraph and a history of earlier paragraphs, a student's next paragraph could compete with the LM-generated continuation of it. This could have had a twofold function. The first aim could have been a training of the deliberate mode in creativity by exploring whether a human evaluator could distinguish between student and LM-produced samples by reconstructing the exact chain of paragraphs generated by the student (with the only cue being the first paragraph that the student wrote). This could have been akin to testing the student's ability to maintain an EB so to speak. The second aim could be a short-term enhancement of divergent thinking in the deliberate mode or a long-term enhancement of the spontaneous mode. Namely, a sort of cognitive stimulation training could thereby be implemented due to the student being exposed to the alternative LM-generated "deepfake science" branch. It is known from cognitive neuroscience, that *"cognitive stimulation via the exposure to ideas of other people is an effective tool in stimulating creativity in group-based creativity techniques"* [191]. Interestingly, the "other" in this case, while not being an EB creator, could be the Type I LM, and the group-based functional unit could be the Type II cyborgnet. The LM in turn could be enhanced by fine-tuning the student's inputs at a later stage. Hence, this educational tool could be called *adversarial cyborgnetic cognitive stimulation*.

- **Epistemic Context Research Papers:** LMs could be used to frequently enhance divergent thinking with regard to the deliberate but also indirectly to the spontaneous creativity mode. Recently, a study demonstrated how GPT-3 can be

utilized as a "multiversal" language model [433], interactively generating branches of fictional counterfactuals to stimulate human creativity in fictional writing. Extending beyond that, scientists could now have combined an LM-aided adversarial cyborgnetic cognitive stimulation with the multiversal approach to GPT-3 to stimulate scientific writing. The fundamental difference with fictional writing would have been that it is the steady application of explanatory criticism by the human combined with adversarially motivated exploration and the possibility of experimental problematization of interest that would have guided the extension of counterfactual nodes. This *multiversal cyborgnetic co-creation* could be further fine-tuned by scientists. Firstly, one could increase the immersion of the LM in the human affective niche via directing its outputs with a slightly altered loss function. Instead of only predicting the next word in a sentence, aesthetic or moral parameters could be for instance considered as well. Secondly, while LMs like GPT-3 are imitative, outcomes perceived as creative are mainly those that exhibit *implausible utility* [501], i.e. utile outcomes with unexpectedly surprising previously underestimated facets. Scientists in their quest for implausible utility, could be inspired by the idea of transdisciplinary cross-pollination effects and insights from research on *cognitive diversity* [362, 432]. Cognitive diversity is related to the differences in information processing and cognitive styles which means it is related to variety with respect to functional features. To fuel intra- and inter-cyborgnetic cognitive diversity with an LM, scientists could be motivated by *composer-audience architectures* [90] from computational creativity [197] utilized to produce humorous outputs by combining an audience model trained on a non-humorous dataset A and a humorous composer model trained on both a different dataset B and the expectations that the pretrained audience model outputs for that dataset [90]. Analogously, scientists could use a dataset from a scientific discipline A and another from a scientific discipline B. A deepfake science LM composer could then learn to surprise a deepfake science LM audience – yielding interesting avenues to augment deliberate and spontaneous creativity but also criticism in the scientists interacting with that *double deepfake science model*. Finally, scientists could harness the knowledge that spontaneous human creativity strongly profits from nocturnal brain processes during sleep [331] to improve the LM's generation of outcomes perceived to stimulate ideas of implausible utility. To this end, they could repeatedly fine-tune the LM on recursively changing text data modified by loosely mimicking e.g. partially sighted evolutionary affective processes of the spontaneous creativity mode [21] extending to synergetic cycles of human sleep [331]. In simpler cases, this could technically include, e.g. targeted *semantic mutations*, *syntactic-semantic crossover* and a form of *semantic noise injection* followed by grammatical auto-reconfiguration at the sentence level. In extensions of such conceptual ideas, scientists could enrich this shifting dataset by letting the LM actively integrate scientific knowledge sampled, e.g. from suit-

able knowledge graphs. Simple active forgetting mechanisms to reduce data size and complexity could e.g. be steered by integrating human preferences via loss functions and/or by integrating human attention during interactions with the LM.

- **Epistemic Context Academic Reviews:** In light of the aforesaid, scientists and educators could practically transform an LM into an interactive *multiversal transdisciplinary deepfake science incubator*. The interesting aspect thereby is that this advanced interactive LM incubator would still *not* be able to understand and create new EBs. This signifies that it could be utilized as a strong baseline offering an enormous amount of material to train the epistemic defenses of reviewers and evaluators against SEA AI attacks. In theory, any conjectured approach to shield peer-review from the new non-EB-like contents of SEA AI attacks must be at least robust against the outputs of that LM incubator at test time. Generally, this could deeply impact and deepen the nature of peer-review. Thereby, the interactive LM incubator could also be utilized for autodidactic purposes and to prepare for red teaming and penetration testing procedures (see Chapter 5). Strikingly, many of the aforementioned could convey humans a sense of *empowerment* emerging from augmentative intra-cyborgnetic feedback loops with LMs. Simultaneously, this could encourage an increased awareness of responsibility on the part of the reviewers and evaluators potentially paired with an altered nature of self-experience via the immensely extended field of affordances for human creativity. In sum, applying generic cyborgnetic defenses to counter SEA AI attacks could at once engender a convergence of moral programming and security research to counter immoral programming.

6.3 Epistemic Meta-Analysis

In our view, in priciple, once a rigorous epistemic elucidation is provided to the general public, humanity as a whole could profit from creativity-fostering *deepfake incubators* via e.g. affordable LM subscriptions that could be available to everyone, such as is the case with access to the internet[1]. Overall, generic cyborgnetic defenses against SEA AI attacks also come with the following inherent caveats: (1) they can be resilient but not immune; (2) they cannot and should not be entirely automated. In summary, in this Chapter 6 and in the last Chapter 5, we pointed to the daunting SEA AI elephant in the room and proposed a non-exhaustive *complementary* solution. The latter could provide cognitively diverse incentives for epistemic security, epistemically-sensitive AI design and also for moral programming for which Wernaart [532] recently set forth a future-oriented road map. We conclude that the international meta-cyborgnet of multiversal scientists may be latently capable of building up *resiliency* against SEA AI attacks. In this vein, may the

[1]Obvious limitations could e.g. be the need to address emerging plagiarism issues [155].

elephant rest in peace. In Chapter 7, we apply the epistemological grounding from this chapter to *virtual reality* (VR). We analyze how one could utilize VR as immersive testbed for a VR-deepfake-aided *epistemic security training* and how present-day AI could act as a catalyst facilitating an epistemically-sensitive *threat modelling* – both for VR and real world environments.

Chapter 7

VR, Deepfakes, Epistemic Security and New Explanatory Blockchains

This chapter is based on a slightly modified form of the publication: N. Aliman and L. Kester. VR, Deepfakes and Epistemic Security. In *2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 93-98. IEEE, 2022. 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.

7.1 Motivation

In practice, while deepfake technology has already been abused for impersonation and cybercrime [486, 437], sextortion and non-consensual voyeurism [210], disinformation and espionage [9, 133], deepfakes in VR [78] may add depth to existing threat vectors next to offering a novel field of affordances for malicious actors – from synthetic non-consensual *VR deepfakes* [121] to immersive disinformation schemes [512] that could even be extended to educational or scientific settings (see also Chapter 3 and 4). Overall, at first sight, it seems that epistemic security considerations caution us against *under*estimating present-day AI and VR when it comes to answering the following question: *do the use and exploit of specific AI and VR technologies risk to harm our own processes of knowledge creation and reasoning?* However, at the same time, when examining the issue from a cybernetic [32] perspective, the following line of reasoning could arise when considering the cybernetic law of requisite variety (illustrated in Figure 7.1) which states that "only variety can destroy variety" [32]. Firstly, when a regulator is faced with adversarial disturbances, one has the following available strategies: 1) reduce the variety of the disturbances (i.e. reduce their degrees of freedom) and 2) augment the variety of the

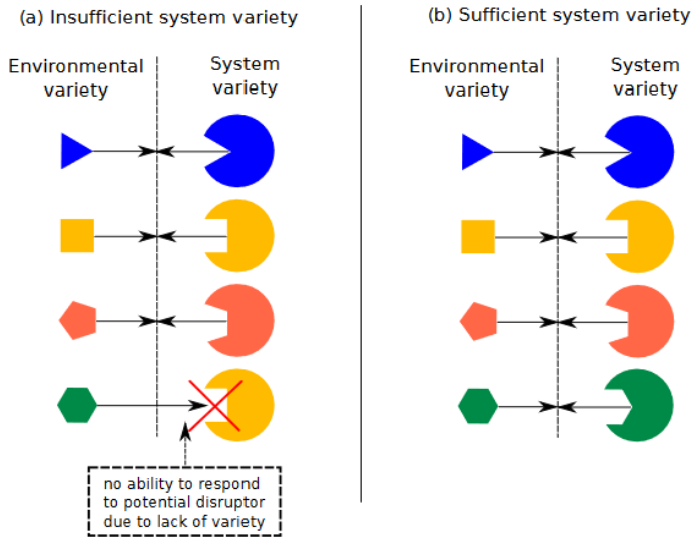

Figure 7.1: Intuitive illustration for the law of requisite variety. Taken from [378].

regulator (i.e. improve its informedness) [22]. Secondly, when practically applied to the contemporary intersection of VR, deepfakes and epistemic security, it becomes clear that one may *not* be able to solve all problems with strategy 1) *alone* since one is not in control of malevolent creativity [140] and may hence need to retreat from society in the long run – which is not realistic for most individuals. For this reason, it makes sense to investigate the avenues that strategy 2) could provide. In short, it may be indispensable to perform transdisciplinary research addressing the following question: *how can we design VR and AI systems that would improve our critical thinking and augment our knowledge creation?*

Thus, while one must not underestimate the negative epistemic impacts that present-day AI and VR technology could cause via malicious actors, it is simultaneously vital not to *over*estimate the capabilities of those systems *if* one is able to dynamically improve one's own informedness. We postulate that the latter is possible – among others through AIVR itself in conjunction with knowledge from epistemological philosophy [200, 202] including the new epistemological postulates summarized in the last Chapter 6. Accordingly, Section 7.2 introduces a theoretical basis for two new transdisciplinary research directions: 1) *VR deepfakes* for epistemic security training and 2) *deepfakes* for cyborgnetic creativity augmentation [16, 22]. The focus is set on the *deepfake text* modality which is often neglected in the public debate despite associated risks [22]. Thereafter, Section 7.3 wraps up and comments on co-creation design fictions connecting the two frameworks.

7.2 Theoretical Basis

7.2.1 VR Deepfakes for Epistemic Security Training

Awareness Creation

In the following, we collate different perspectives on why VR may serve as an excellent playground to raise awareness for epistemic security in the deepfake era to begin with. Firstly, multiple studies corroborated that VR can successfully create security awareness in diverse application areas ranging from cybersecurity in educational settings [517] to security in critical infrastructure facilities [145, 150]. Secondly, in the context of cybersecurity training, gamified VR settings [209] were conducive to enhanced learning experiences (in comparison to traditional lessons) [521]. Thirdly, in the nascent field of immersive journalism [401, 500], VR has been described to be able to offer a unique grasp on a given situation by *"transferring people's sensation of place to a space where a credible action is taking place that they perceive as really happening, and where, most importantly, it is their very body involved in this action"* [153]. Furthermore, De la Peña (who has also been called the "godmother of VR" [300]) adds that VR facilitates a *"profoundly different way to experience the news, and therefore ultimately to understand it in a way that is otherwise impossible, without really being there"* [153]. On the whole, it is easily conceivable that a coalescence of the following elements may enable synergetic effects fostering an awareness of the deepfake-related epistemic threat landscape: 1) VR environment, 2) gamification and 3) VR news contents on past real-world deepfake attacks involving malicious actors.

Epistemic Calibration

Once an initial awareness would have been achieved, it may become expedient to facilitate an *explorative* educational VR deepfake setting. By way of example, one could implement a VR platform where individuals can experience state-of-the-art deepfake text capabilities by interacting with a set of virtual avatars some of which are driven by present-day language AI [177, 245] and some of which are embodiments of human users. In this way, people could actively improve and test their knowledge on when one must avoid to *over*estimate present-day language AI and in which contexts on must conversely not *under*estimate its abilities. It is cogitable that with improving abilities, there may be no fundamental limit to the accuracy with which present-day language AI may be able to imitate all linguistic outputs *that are imitable* [77]. However, recently, it has been predicted that it would stay *impossible* for any imitation-based AI to ever be able to create *new* so-called explanatory blockchains (EBs) [16, 22]. Hence, as a scientifically interesting side-effect, the proposed educational VR deepfake setting would also allow the rigorous

testing of this falsifiable impossibility statement via new to be created EBs. In brief, the format of new EBs corresponds to (or can be easily converted to) the robust format that was underlying the chains of explanations inherently instantiated in humanity's best tested scientific theories (including formulations of laws of nature), best patent applications and best philosphical frameworks – at a time when those were new.

Probing of Defenses in Blind Settings

Regarding defense strategies against deepfake attacks, one often encounters short-term-focused source-based schemes such as authentication, identity corroboration and deepfake detection. However, a *complementary* long-term approach is required in which one foregrounds the *content*. The reason being that source-based methods are relying on trust that can be eroded or on security mechanisms that can be bypassed by adversaries informed of the defenses in place. Also, source-based defenses risk to unintentionally worsen a stigmatization [294] of human statistical outliers (think e.g. of individuals with physical disabilities in visual deepfake contexts or neurodivergent individuals such as autistic people with idiosyncratic writing styles in deepfake text contexts). For this, interactive educational VR deepfakes may provide invaluable avenues by virtue of the dynamically adjustable blind setting that VR offers – where one could calibrate potentially deceptive source-related parameters on social cues and linguistic style linked to cognitive biases already exploited in social engineering [274]. While CAPTCHAs [221, 313, 333] are in principle content-based, there are known to yield incessant cat-and-mouse-games which is why it makes sense to consider novel more robust alternatives. For instance, to counteract deepfake science attacks based on deepfake text, it has been suggested that since near-future language AI may be able to imitate any imitable linguistic output, science must shift the strategy *away* from deepfake text detection attempts [22] and that instead of trying to identify the source of a text sample, the goal would now be to identify whether it contains a new EB [16]. The latter would namely corroborate cognitive efforts spent by an agent able to understand explanatory information (EI) – with humans being the only species and only group of entities on Earth able to fulfil this condition.

In this way, using the generation of new EBs as epistemological baseline (see also Chapter 6) improving beyond vrCAPTCHA [333] ideas, one can achieve an *asymmetric* test framework [16, 22] that is qualitatively different from Turing Test schemes. While the identification of a new previously unknown EB in a text corroborates the participation of an entity able to understand EI, the absence of a new EB does *not* signify that the text was deepfake-generated. Firstly, note that the text could still have been generated by a person that was simply not willing to generate a new EB, not ready for it, intentionally doing the contrary and so forth. Secondly, while the presence of a new EB corroborates the cognitive efforts of at least one entity able to understand EI (i.e. a person), it does *not*

signify that this entity did *not* use non-EB-like deepfake material for purposes of creativity stimulation or paraphrasing. In sum, a test based on novel EBs is asymmetric with regard to the test outcomes since a positive test leads to a *homogeneous* group of entities that corroborated their ability to understand EBs (which implies an understanding of EI) and a potentially *heterogeneous* group of entities that could contain both entities that *do* understand EI (such as humans) and entities that do *not* (such as present-day language AI). Overall, when equipped with these asymmetric epistemic premises, a VR deepfake setting could improve the *critical thinking* abilities of both scientists and nonscientists.

7.2.2 Epistemically-Sensitive Deepfake Design

Epistemic Calibration

Before elucidating why one could harness deepfake text to augment human creativity *in the service of epistemic security* both in real-world environements and in VR, we briefly clarify the terminology. Firstly, the term "cyborgnet" [16, 22] stands for a generic, substrate-independent and hybrid functional unit. A cyborgnet is much more general than and *not* to be confused with the term "cyborg" (i.e. while all cyborgs exist in cyborgnets, the reverse does *not* hold). The minimum requirement for a cyborgnet is a directed graph where explanatory narratives combine: at least one entity that *does* understand EI (such as e.g. humans) and at least one entity that does *not* (such as e.g. present-day language AI, chairs, thoughts, stone tools, fishes and so forth). Crucially, because language itself can be considered to be a technological tool [173], a human already existed within a cyborgnet since the dawn of language. Secondly, since couplings of present-day language AI and humans qualify as instances of cyborgnets, it follows that anthropic creativity augmentation using this AI represents one form of *cyborgnetic creativity augmentation* [16, 22]. In the following, we elucidate why cyborgnetic creativity augmentation could contribute to epistemic calibration in the context of deepfake attacks both in real-world environments and in VR settings.

From a psychological perspective, it is apparent that against the already existing background of disinformation and so-called "fake news" [122] phenomena, advancements in deepfake technology paired with a stronger harm intensity [240] of technically feasible deepfake attacks [250] could have the potential to exacerbate pre-existing human epistemic fear constructions. The latter may be reflected in contemporary usages of expressions such as e.g. "epistemic anarchy" [280], "post-epistemic world" [260] and "post-truth era" [307]. Furthermore, the deepfake threat landscape engendered *automated disconcertion* (see Chapter 2) – the epistemic confusion that arises merely by the possibility of malicious deepfakes. In the light of the aforesaid, it is easily conceivable that adversaries could exploit the contemporary fragile epistemic ecosystem and instrumental-

ize automated disconcertion. In the main, a security-aware defense strategy must thus proactively counteract any unnecessary *over*estimation of deepfake capabilities to avoid fostering adversarial success. As already adumbrated in Section 7.1, if as a defender one cannot reduce adversarial disturbances, an alternative strategy is to augment one's own variety (i.e. improve one's own informedness). In sum, for a better epistemic calibration concerning the deepfake technology owned by malicious actors, one can improve one's own informedness. Strikingly, one possible way to achieve the latter in AI and VR would be via a cyborgnetic creativity augmentation scheme harnessing *deepfake technology itself*.

Before illustrating a deepfake-text-aided cyborgnetic creativity augmentation, we compactly recapitulate a background from epistemological philosophy needed to get a better model on how present-day language AI *could* but also *could not* augment our knowledge in the first place. As stated by Popper, it is impossible to scientifically predict the future of knowledge creation [410]. This is why an imitative AI could not learn from historical data how to create new previously unknown EBs (as elaborated in Section 7.2.1). Metaphorically speaking, Popper would have agreed that an imitation-based AI able to reliably predict any future new EB would be an *epistemic perpetuum mobile*. Beyond that, as reinvigorated in a regimentation of Popperian critical rationalism by Frederick [200, 202], the epistemic aim of science is to achieve *better* explanations [200] – and *not* "truer" ones for lack of a direct, verbalizable access to truth from the stance of a knowing entity[1] [200]. (The latter was already implied by Kant [282] when considering the unknowable "Ding-an-sich".) To translate it to the vocabulary from Section 7.2.1: our epistemic aim can be to achieve better new EBs. Ergo, in cyborgnetic creativity augmentation with present-day language AI, while the latter *cannot* generate new EBs, it *can* stimulate human creativity with new non-EB-like EI and any other non-EB-like linguistic output.

Multiversal Threat Modelling

On the one hand, in areas such as cybersecurity and security for machine learning, it is indispensable to perform a so-called *threat modelling* [101], a specification of goals, capabilities and knowledge exhibited by a given adversary. On the other hand, in human-computer-interaction (HCI), so-called design fictions enable *"HCI and design researchers to co-create, explore and speculate the future"* [6]. To augment threat modelling and increase its graspability in security and safety domains, one can craft design fictions grounded in threat models (see Chapter 2) – which may also be relevant for epistemic security in the context of deepfake attacks. In recent years, the concept of co-creation design fictions [406] has been applied to various domains including AI safety [262], AIVR

[1] On a *deflationary* account of truth [87] that does *not* equate it with consensus as often inaptly intrinsically done in colloquial languge including in AI contexts [172], we do *not* inhabit a post-truth era (see also Chapter 3). Moreover, we do *not* inhabit a post-falsification era.

safety and VR security [503]. Generally, as stated in Chapter 4, design fictions *"can be used for technological future projections by experts in the form of for example, narratives or construed prototypes that can be represented in text, audio or video formats but also in VR environments"*. One way to augment human defenders in threat modelling including design fictions would be to use present-day language AI to generate creativity-stimulating linguistic outputs extending the space of ideas for threat models [16] (see also Chapter 5). (Thereby, as mentioned earlier, it is impossible to implement an oracle able to predict the future creation of new EBs. Hence, neither design-fiction-augmented threat models taken alone nor present-day language AI utilized in cyborgnetic creativity augmentation schemes could fulfil the role of oracles.) As we explain in the following paragraph, in the deepfake era, a cyborgnetic creativity augmentation using *deepfake text* may even need to be integrated in a defender's toolbox *per default*.

Firstly, large language models have been described to be able to enhance human creativity [325, 557, 561] especially by generating counterfactual text samples that unfold a "multiversal" [433] approach. In this sense, for any subfield engaging in counterfactual risk analyses [539], one could harness present-day language AI trained on historical samples of relevance for the subfield in question in order to augment threat modelling by defenders [16] – which could profit from looking around corners and propagating through mental barriers by contemplating non-EB-like deepfake text counterfactuals. The latter may be of interest for cyborgnetic creativity augmentation including "multiversal cyborgnetic co-creation" schemes [22] for science in general. Secondly, one must take into account that malicious actors could use deepfake text to fabricate misleading cyber threat intelligence [425] able to deceive both humans and present-day AI pipelines, to fabricate misguiding non-EB-like explanations for world events linked to fictional synthetic histories [260] or to perform non-EB-like deepfake science attacks [16]. Importantly, the generation of such powerful (but still non-EB-like) counterfactual material implied in those schemes already signals the possibility for such an attacker to harness cyborgnetic creativity augmentation for own malicious goals. Thirdly, due to the latter, a responsible counterfactual risk analysis may need a design-fiction-augmented *multiversal threat modelling* using deepfake text. This could foster multiversal epistemic security both for the real-world and in VR.

7.3 Conclusion and Future Work

Given that *unethical* actors could harness deepfakes and VR deepfakes to harm human epistemic processes [94, 232, 336], novel robust and tailored epistemic defense strategies are required *from the onset on* – and not in hindsight. As a *complementary* contribution to ongoing efforts along those lines, we presented two new AIVR research directions on how

defenders and developers could design VR and AI technology that would instead improve human critical thinking and augment human knowledge creation. Thereby, for illustrative purposes, we focused on the often underestimated *deepfake text* modality which is linked to neglected attack vectors including instances of deepfake science attacks [16, 22]. Firstly, we explicated how and why an amalgamation of deepfakes, gamified VR environments and immersive news on past deepfake attacks could be used for a new form of *epistemic security training* in VR able to enhance human critical thinking. Secondly, we expounded how and why AI could serve as catalyst within a deepfake-text-aided cyborgnetic creativity augmentation allowing an epistemically-sensitive *multiversal threat modelling* enriched by co-creation design fictions.

In sum, while applying knowledge from cybernetics and epistemological philosophy to modern deepfake issues, this chapter illustrates how *AIVR safety* suddenly becomes unified with *AIVR ethics* whilst crafting defenses against epistemic security threats. The latter corroborates the ability of VR frameworks to serve as experiential testbed for ethical decision-making [12] in socio-psycho-technological contexts. In future work, one could perform co-creation design fictions enabling people to explore a counterfactual future in which deepfakes are more profoundly integrated in societal structures. It could for instance be framed as *gamified multiversal threat modelling for epistemic security* in social VR (e.g. as *fictive* deepfake-text-augmented poll on AI-as-a-service schemes [335] for future *synthetic* societal functions from "AI co-workers" over "deepfake psychologists" to "VR politicians") – possibly with insights for real-world defenses against deepfake attacks. Conceivably, in habitually *non*-explanatory-blockchain-like settings, severe problems of indistinguishability could arise. However, as long as it is not made problematic by experiment and provisionally refuted by a better new theory, explanatory-information-understanding entities such as humans can purposefully use *new* (i.e. previously unknown) *explanatory blockchains* as an albeit asymmetric epistemic shield when needed. This asymmetric shield can boost the robustness of science against deepfake science attacks. In brief, while one could employ deepfakes to harm the epistemic processes of an *unprepared* society, deepfakes are *not* an epistemic perpetuum mobile without remedy.

7.4 Epistemic Meta-Analysis

This VR-focused chapter naturally unified strategies for epistemic security and epistemically-sensitive AIVR design. One recurring key point is that an epistemic perpetuum mobile is impossible: creating new EBs comes at the cost of a harder Type-*II-only* process of *understanding* which requiring cognitive efforts. In the next Chapter 8, assuming *that* this is the case, we first analyze the implications thereof for the "meaningful" control of Type *I* AI. Then, in Chapter 9, we explain *why* new EBs could be epistemically special.

Chapter 8

From OODA-Loop To COOCA-Loop

8.1 Motivation

Given the theoretical background from Chapter 6 stating that Type I AI (i.e. including all present-day so-called intelligent systems) can neither understand EBs nor create new ones, one can anticipate a *cyborgnetic comprehension bottleneck* that could arise in uninformed attempts to control it. Indeed, the cyborgnetic comprehension bottleneck can be understood as a consequence of the asymmetry between the ability to create information of the form x and to understand that information x which was mentioned earlier in Chapter 1. In this context, one can start by examining the epistemic problems emerging in the extreme case of an intelligent system instantiating a classical OODA (Observe, Orient, Direct, Act) loop as a fully automated, i.e. end-to-end-Type-I pipeline. We remark that in high-risk contexts and strategically complex domains, a reasoning via EBs may (and one could even state should) play a particularly important role. However, it is now apparent that if the AI goal framework for the Type-I-OODA-loop pre-determined by humans would have been developed based on EB-based reasoning, the AI would *not* be able to enact its meaning in new contexts. The latter is given since it is considered to be impossible for a Type I AI to create new EBs. This represents a strong limitation to any conception of run-time "adaptivity" in EB-based decision-making including e.g. EB-based *moral reasons* [105]. Note that a heterogeneous mixed scenario in which some functions of the OODA loop are delegated to Type II entities but there exists (at least) one Type-I-*only* function does *not* solve the comprehension bottleneck problem as no novel EB-based message passing can be reliably implemented. Then, at first sight, consistent with the arguments presented in this book, it may seem recommendable to specify the requirement for high-risk contexts that *each single function* of an OODA-loop must be *cyborgnetic*. (A cyborgnet as a whole is always of Type II since it contains at least one Type II entity. Crucially, note also that a cyborgnet need *not* include any Type I AI since

e.g. an individual human inherently lives in language and already fulfils the definition of a cyborgnet.) However, in the following, we explain why strictly speaking, for epistemic reasons, one would then need to extend beyond the notion of an OODA-loop.

8.2 COOCA-Loop Meta-Paradigm

Firstly, an OODA-loop could *not* epistemically be cyborgnetic because no reasoning in Type II entities such as humans begins by induction [411]. In short, strictly speaking, no conscious OODA-loop actually starts with an observation. Instead, as already hinted by Popper [411], there must first be *a point of view* from which we actively sample the world – by what perception is inherently conjectural i.e. theory-laden. For this reason, a cyborgnetic OODA-loop would only stay an oxymoron. Thus, a first step is to explicitly add the following function: Conjecture (abbreviated with C in the following). Secondly, we explain why it is sensical to transform the Decide (D) function into a novel Co-create (C) function. Classically, in the AI field, decision-making is associated with a known set of options from which one has to choose. However, due to their own creativity capabilities and conscious choices, Type II entities can decide to create new options or even to destroy old ones. In brief, the space of options is strongly dependent on Type II creativity since it can ultimately contain the creation of new EBs (which can even include a revaluation of values [157]) for which it is impossible to reliably predict them ahead of time. Even where humans pre-specified an intention to throw dices to resolve unconclusive issues, *"uncertain humans equipped with some dice at the time of moral decision making could throw that dice but could also unexpectedly (co-)create novel as yet unknown solutions on how to solve the problem"* [22] – something present-day "AIs" cannot.

Thirdly, one can now integrate the generic concept of AI-based cyborgnetic creativity augmentation exemplified in Chapter 6 and 7. In theory, this now becomes possible at the level of *each individual function* since each one is itself cyborgnetic. In general, to omit opportunities for creativity augmentation where adversaries practice it could be especially detrimental. It thus seems recommendable to implement it where practically feasible. To sum up, we just explained why for epistemic reasons, one requires the novel *meta*-paradigm of a cyborgnetic COOCA (Conjecture, Observe, Orient, Co-create, Act) loop for responsible AI design. Beyond that, morality could be (and ideally should be) EB-based and it is impossible for an automated, i.e. end-to-end-Type-I-only pipeline to adaptively create novel EBs on-the-fly post-deployment. Hence, in the same way as we stated that an epistemic perpetuum mobile is impossible (see Chapter 7), one can conclude that a *moral perpetuum mobile* is impossible too. While one can use Type I AI for moral augmentation *within* a cyborgnet, one will *not* be able to reliably outsource the cognitively demanding task of creating new EBs solving moral problems to Type I AI.

Some past approaches to responsible AI design are already intrinsically compatible with the *generic* COOCA-loop meta-paradigm and appear epistemically permissible as follows:

- ***Inter*-function-level:** Since *each single function* must be cyborgnetic, there must be at least one Type II entity in *each* function to account for the possibility of EB-based communication *between* the functions. This is instantiated by some *human-in-the-loop* approaches. Also, recall that a Type I AI in a function is *not* obligatory.
- ***Intra*-function-level:** While each high-level function must be cyborgnetic, there is room for improvement *within* an individual function. There, *where feasible*, one can improve speed, scale and scope by harnessing *local* Type-I-OODA loops. This allows any of the three paradigms encapsulated *locally within* the cyborgnet: human-before-the-loop, locally unsupervised loop and human-in-the-loop.

8.3 Local Intra-Function Encapsulation of Type I AI

Against the background of the explanations from the last Section 8.2, we revise the role of previous transdisciplinary frameworks for meaningful Type I AI control – especially in high-risk contexts. A recent example is the *meta*-ethical and *non*-normative *augmented utilitarianism* (AU) which was designed to serve as a suitably structured but empty Type I AI goal scaffold left blank *"in which moral authorities (especially users or society but also designers in default settings) fill in flexible updatable and machine-readable heuristic moral models"* [26]. The authors described that *"AU targets what one could conceive of as a possible smallest heuristic moral superset (SHMS) capturing the plurality of candidate ethical frameworks available in practice for moral programming"* [26]. This SHMS was specified to be currently representable as encompassing morality-relevant parameters on: 1) perceiver, 2) agent, 3) action, and 4) patient. Retrospectively analyzed, we argue that due to the possibility of human preferences for (and even recommendability of) dynamic EB-based reasoning applicable to morality, Type I AI cannot reliably solve the so-called *moral chunking* problem – the moral chunks could unpredictably be new yet unknown EBs. For more clarity, consider that one could first attempt to extend the SHMS to now contain *at least* the following five elements: 1) perceiver, 2) agent, 3) action, 4) patient *and* 5) new EB(s). However, not only is it impossible for Type I AI to identify new yet unknown EBs themselves but in addition, the creation of new EBs is even itself able to modulate *action* and/or *perception* (including the categorization of *agent* and *patient*) via *affective realism* [50, 204] (see also Chapter 3). For example, in light of a new EB, a human faced with a context deemed to be a morally-relevant situation could suddenly evaluate that situation entirely differently – as it could literally be analyzed "through a different lens". In addition, a situation previously perceived as morally-relevant could even now be constructed as being neutral and vice versa.

It seems that to merely state the feasibility of heuristic moral models to control Type-I-only-loops risks to nourish the misconception that present-day AI is assumed to understand the moral reasoning that generated those heuristics. While this is *not* the case in AU (which describes AU-based heuristic moral models merely as *"ephemeral approximative shadows of morality to be necessarily updated with time"*), one needs to more actively counteract the epistemic overestimation of Type I AI. Also, given the non-normative nature of previously described AU moral models, one criticism could be the risk for moral relativism [61]. On the whole, it becomes clear why for a responsible and simultaneously efficient epistemically-sensitive AI design, one would now have to encapsulate a framework initially designed for the governance of Type I "OODA-loops" such as AU *within* an individual function of a COOCA-loop. In this way, there is a transparency about the epistemic capacity of e.g. an AU-governed Type I AI encapsulated at the intra-function level of a COOCA-loop. Namely, because this Type I AI is not able to understand the most complex epistemic artefacts that a COOCA-loop could transmit between its functions, it can only complement but never substitute a single function. Moroever, because each single function is necessarily *cyborgnetic* (i.e. inherently of Type II) and EB-based morality is even recommended, the moral agency of cyborgnetic entities (i.e. here humans) is emphasized. The latter counters moral relativism in that the very design of the COOCA-loop requires that the participating Type II entities are already cyborgnetically rational[1] – which simply signifies that those Type II entities are aware of being *able* to create better new EBs *if willing to* and when needed. Interestingly, for some, the creation of new EBs itself could then simultaneously serve as both norm and value. In this sense, nowadays, one could regard new EBs as being among the hard*est*-to-vary novel unpredictable affordances that living entities can create. But it is precisely this form of originality that cannot be modelled by AU-governed Type I AI and whose contents can neither be specified quantitatively nor qualitatively ahead of time. In this vein, as stated by Bohm [74] it holds that *"[...] to define originality would in itself be a contradiction, since whatever action can be defined in this way must evidently henceforth be unoriginal."*

[1]Note that in contrast to dualistic accounts of rationality that attempt to divorce human rationality from affect, cyborgnetic accounts of rationality acknowledge that the latter is impossible since as described in Chapter 3, continuous affect (but *not* necessarily discrete constructions such as arbitrary emotion categories [50]) is inseparably entangled with the experience of embodied consciousness [50, 255]. What is more, the cyborgnetic notion of rationality sketched here necessarily and moreover even *explicitly* includes affect in a subtle way. Namely, this is already the case via the criteria to identify "better" EBs in the first place. As described by Popper, justifications are logically invalid [58] while the therefore necessarily non-justifiable updatable criteria for better EBs include affective components with varying constellations of arousal and valence. The latter is e.g. more pronounced when it comes to the preference for EBs that are "more aesthetically appealing" (which often plays an important role in e.g. cosmological theories of physics). Obviously, also the updatable criteria for the novelty of a given EB cannot be separated from affect. Even when harnessing a Type I AI to "detect" novelty, the latter is done on the basis of preceding updatable cognitive-affective constructions of what novelty signifies (see also e.g. Chapter 6 for EB-novelty criteria adapted to the deepfake era).

8.4 Global Inter-Function-Level Epistemic Security

In this section, we briefly speak to the advantage of utilizing COOCA-loops instead of those instantiations of "OODA-loops" that would formally *not* correspond to a COOCA-loop (i.e. including but not limited to Type-I-only loops). In the current adversarial AI field, a known technique is the development of automatable so-called substitute models [142, 264] that are utilized to simulate a more or less opaque victim AI model that an attacker intends to target subsequently. Given that there is no limitation to the accuracy with which one could attempt to forge the creation of new non-EB-like information, any Type-I-only-function could be in theory automatedly simulated by an adversary. While it might be tempting to assume a black-box setting if there is a secrecy of software, one must avoid security-by-obscurity pitfalls [508] and it is more prudent to realize that ultimately a grey-box or even a *white-box* setting can be achieved in case of preparatory adversarial attacks e.g. via physical stealing of AI equipment, model stealing via an application interface to a similar AI model or more classically via data theft in a cyberattack – all of which can reveal the internal specifications of the deployed Type I AI. By contrast, the worst-case scenario for a cyborgnetic function (i.e. here including humans) would lead to a grey-box setting which can be achieved via information gathering harnessed to gain crucial personal information. In sum, in worst-case-scenarios, a Type-I-only-loop could yield a white box setting while a COOCA-loop as a whole (with or without Type I AI included within individual functions) can at worst become a *grey-box* setting to an adversary – because Type *II* entities can unpredictably create new yet unknown EBs that neither Type I nor even Type II entities could reliably predict ahead of time.

8.5 Epistemic Meta-Analysis

8.5.1 Relevance for AI-Related Epistemic Security Strategies

In the last Section 8.4, we adumbrated that a COOCA-loop offers comparatively more secrecy than an alternative meta-paradigm where one or more Type-I-only function(s) would be allowed. For instance, it is conceivable that a COOCA-loop could profit from potential cryptographic bonuses that deepfake artefacts[2] such as language-AI-generated text could offer. It could confer the ability to conceal own new EBs by randomly intermingling those with specifically crafted counterfactual deepfake text [16] building on and extending highly interesting earlier deterrence strategies against intellectual property theft [1]. Via a COOCA-loop, one obtains the possibility of *secrecy-by-epistemology*.

[2]Other examplary use cases could be so-called honey tokens [567] generated on the basis of deepfake code [211] and even future "honey social VR rooms" [16] to distract potential Type-I-pipelines harnessed to preferably effortlessly extract intellectual property instantiating secret new EBs.

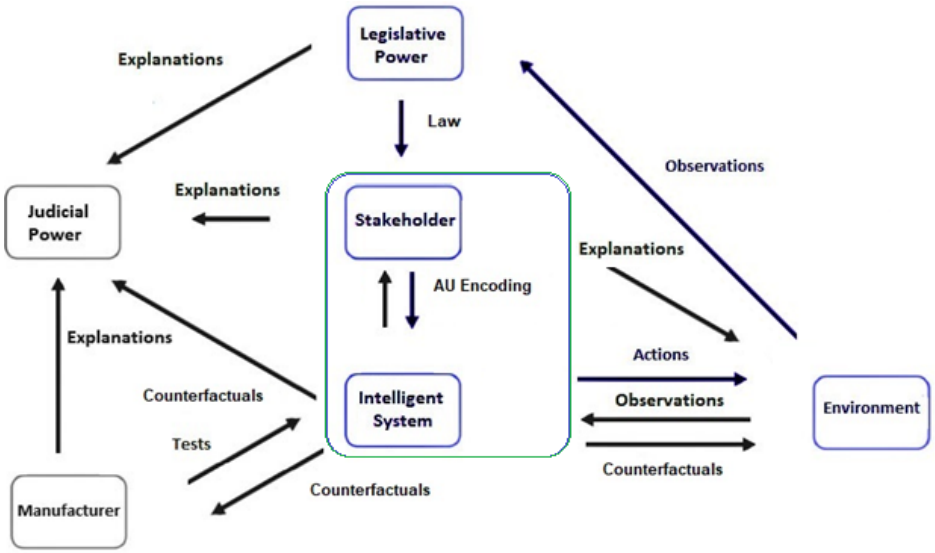

Figure 8.1: Simplified illustration for a snapshot of epistemic processes in a larger cyborgnetic socio-technological feedback-loop under the COOCA-loop meta-paradigm. The Type *I* intelligent system is locally *encapsulated* within the *cyborgnetic* function of the unpredictable Type *II* stakeholder. The latter has the possibility to adaptively generate new yet unknwon better EBs when required. Adapted from[24] and modified.

8.5.2 Relevance for Epistemically-Sensitive AI Design

As elucidated in this chapter, the concept of an OODA-loop needs to be replaced by the meta-paradigm of the COOCA-loop. In Section 8.3, we explained why in high-risk contexts, for a responsible epistemically-sensitive AI design, Type-I-only-pipelines must stay locally *encapsulated* at the *intra*-function level within an individual function of a more global COOCA-loop. Using the example of AU-governed Type-I-AI-loops, in no case should one utilize them to maintain artifical moral filter bubbles facilitating an epistemic stagnation in outdated assumptions. Instead, one could efficiently use AU-governed Type-I-AI-loops *locally within* a cyborgnetic function of the COOCA-loop e.g. as follows [26]: 1) to augment anthropic moral reasoning; if desired with the AU-encoded augmented utility function option, 2) to *virtually* test a socio-technological feedback-loop using simulation environments, 3) to exploit counterfactuals and craft adversarial AU-based augmented utility functions for self-education or creative exploration, 4) to virtually simulate deepfake attacks to enhance epistemic security in various domains. To wrap up, Figure 8.1 displays a simplified illustration of a snapshot from an individual cyborgnetic function located within a larger socio-technological feedback-loop respecting the COOCA-loop meta-paradigm. Future work could deepen this research including the legal implications of the required EB-based accountability in the deepfake era. However, in Chapter 9, we finally cover the twofold open question: 1) *why* is there a qualitative epistemic gap between Type *I* and Type *II* entities and 2) can one build Type *II* AI?

Chapter 9

The Cynet Butterfly Effect

9.1 Motivation

9.1.1 Epistemic Security Paradigms: AS versus EC

In Chapter 2 we contrasted two different *epistemically-relevant* paradigms in AI safety which differ fundamentally in the long-term policies they imply. While the first paradigm is called *artificial stupidity* (AS), we termed the second one *eternal creativity* (EC). To recapitulate, AS postulates that AI ability needs to be upper-bounded by human performance since it risks to otherwise become uncontrollable. In short, AS suggested that AI will need to be made *artificially stupid* [499] for AI safety reasons. By contrast, refining it with the notions gradually developed and improved in the last chapters and deepened in this very Chapter 9, the renewed EC paradigm which is instated in this book recommends the *augmentation* of *creativity* within a *cyborgnet* following the *COOCA-loop* meta-format from Chapter 8. This includes but is not restricted to processes of artificially augmenting human creativity and processes of augmenting the creativity of present-day AI. These fundamental epistemic differences are reflected in the contrast between long-term AS guidelines versus long-term EC guidelines – with significant implications for *epistemic security* strategies (see Chapter 2). While AS yielded *intelligence*-focused, *restriction*-based and substrate-*dependent* long-term guidelines, EC proposed *EB-creativity*-focused, cyborgnetic-creativity-*augmentation*-fostering and substrate-*independent* long-term strategies. The latter is not surprising given that the epistemic premises of AS and EC are fundamentally different.

Generally, in AS, a superintelligence is understood as an intellect exceeding human cognitive performance in *"[...] virtually all domains of interest"* [79]. (Thereby, as described in Chapter 2, AS distinguishes between three types of superintelligence: speed, collective and quality superintelligence.) However, in this form, the paradigm is *not* yet amenable

to *experimental* problematization since the set of all domains of interest is not explicitly specified. By contrast, in EC, the domain of interest is explicitly the task of reliably creating new better EBs – which is postulated to be fundamentally impossible for Type I entities but possible for hereto willing Type II entities. In brief, the formulation of EC is comparatively simpler, clearer and more risky – which is easing (instead of risking to hamper) experimental problematization and critical *scientific* analysis. Indeed, a monolithical focus on too carefully formulated statements that cannot be directly made problematic by experiment and allow manoeuvres that risk to artificially maintain those staments alive risks to engender an epistemic stagnation where dynamic epistemic updates would be necessary instead. In sum, to provide a more robust grounding for *AI-related epistemic-security strategies* in the deepfake era, this book instates the EC paradigm. Thereby, the role of this chapter is to compactly collate a transdisciplinary set of *explanatory* frameworks from various scientific areas to elucidate *why* a qualitative difference between Type I and Type II entities is conjectured in EC.

9.1.2 Fundamental Difficulty of Type II AI Design

Moreover, this analysis simultaneously provides an answer to a pertinent question that may be relevant for *epistemically-sensitive AI design*: (how) could one implement a Type *II* AI? The AS paradigm assumes that at least a speed or a collective superintelligence could be implemented artificially by humans within a timeframe that is as imminent that it needs to be integrated in the AI safety policies of the present decade. Thus, applying the cyborgnetic terminology to it, it becomes clear that AS assumes that the artificial implementation of a Type II AI (which both the hypothetical speed and collective superintelligences of AS would be) is not only possible but also imminent. Beyond that, upon closer analysis it also becomes clear that in addition, when applying a cyborgnetic terminology, the AS paradigm implies that a third type of system that would be qualitatively better than a Type II entity is possible since a *qualitative* superintelligence is assumed to be able to be *"at least as fast as a human mind and vastly qualitatively smarter"* [79]. From a cyborgnetic perspective, this would imply a bizarre sort of "Type III" entity. However, given that again no concrete experimental problematization is provided (see Section 9.6), there is no "artificial superintelligence" yet even according to AS definitions and no explanation on how a qualitative (and not merely quantitative) superintelligence could be reliably implemented by qualitatively "inferior" entities is known, it seems that currently, such a "Type III" cluster is superfluous and does not represent a genuine problem to consider [16]. To sum up, EC assumes that quantitative measurements of intelligence are not the top regulatory priority, instead it foregrounds the fundamental distinction between Type-I-accessible and Type-*II-only*-accessible creativity – two qualitatively different creativity categories that coalesce in cyborgnets and synergetically form *cyborgnetic creativity*. In the following Section 9.2, we elucidate how

Focus	Complex	Living	Conscious	Cyborgnetic
Ex.	e.g. in [470], [349], [525], [304], [52], [65]	e.g. in [135], [438], [188], [68], [189], [68]	e.g. in [538], [152], [293], [377], [375], [230], [51]	e.g. in [130],[183], [399] [128], [16], [17]
D	neuroscience [470]; chaos theory [349]; complex systems theory [525]; biology [65]	physics [135]; biology [438]; biophysics [188]; biorobotics [68]	psychology [538]; enactivism [152]; biosemiotics [377]; physics [293]; biology [375]; neuroscience [230]	psychology [130]; physics [183, 399]; philosophy [128]; cyborgnetics [16]
CyT	complex but not necessarily living entity	living but not necessarily conscious entity	conscious but not necessarily Type II entity	necessarily a Type II entity since cyborgnetic

Table 9.1: Simplified collection of exemplary high-level explanatory focuses for *non*-reductionist frameworks. Ex. denotes exemplary studies. D specifies exemplary disciplines or meta-frameworks in which those studies are embedded. CyT corresponds to a short comment on the focus as seen trough a cyborgnetic theoretical lens. As reflected in the CyT row, the scope of the explanatory frameworks expands from left to right.

resonating with modern explanations from biology, physics, neuroscience and philosophy, EC implies that Type II entities like humans cannot be reliably modelled by Type I entities like present-day AI. In short, EC contradicts the view that humans are reducible to a Turing Machine [44] (being a Type I entity) which we call *the reductionist approach*. We explain why similarly to AS, a Type II AI is in theory possible under EC but why in contrast to AS, EC concludes that: 1) to build a Type II AI is a task of *universal* difficulty involving the to be introduced *cynet butterfly effect* (see Section 9.4), 2) "Type III" AI is an unnecessary worry since it is scientifically impossible as explained in Section 9.7.

9.2 From Complex Dynamical Systems to Dynamic Universal Creativity

9.2.1 Non-Reductionist Explanatory Frameworks

Table 9.1 collates a non-exhaustive structured set of exemplary *non*-reductionist explanatory frameworks that could explicate why Type-II-ness and inherently the creation of new EBs could represent epistemically special phenomena – as conjectured in EC. We group

those frameworks in four clusters with four attributes reflecting their implicit or explicit focus: 1) complex, 2) living, 3) conscious and 4) cyborgnetic. In the first cluster "complex", non-reductionist explanatory frameworks elucidate why complex systems cannot be reduced to linearly evolving systems. In the second cluster labelled with "living", the frameworks emphasize that the degrees of freedom exhibited by the emergent dynamics of living entities cannot be reduced to inert building blocks. Explanatory frameworks from the third cluster "conscious" stress that consciousness is able to shape biological evolution and that conscious beings cannot be reduced to biological functions. Finally, in the fourth cluster "cyborgnetic", corresponding research explicates why the study of Type II entities (i.e. including but not limited to humans) transcends everything else and *irreducibly* gains a *universal* scope. On the whole, the latter may not appear supprising anymore since when synthesizing the exemplary frameworks illustrated in Table 9.1, one can extract that Type II entities such as humans exhibit a muli-level irreducibility by virtue of simultaneously being complex, living, conscious *and* cyborgnetic. Next, we briefly illustrate each cluster with exemplary explanations from the studied literature.

Complex Systems

Firstly, the cluster "complex" cautions scientists against underestimating the intricacy of modelling Type II entities like humans since human cognition and behavior can exhibit the peculiarities of *complex dynamical systems* [304] (which can range from complex non-living Type I physical systems such as the weather [340] over the immune system [65] to human group behavior) for which it holds that *"the whole is greater than the sum of its parts"* [304]. For instance, Barrett [52] states that reductionism is even impossible in practical psychological research as *"[...] a living organism is not an assemblage of separable mechanisms that can be studied bit by bit. Rather, contextual factors that may be weak on their own interact and coordinate in nonlinear ways to powerfully create phenomena that cannot be reduced to any weak factor in isolation. And importantly, it is not possible to manipulate one factor separately and leave the others unaffected"* [52]. Generally, the concept of a complex dynamic system is often linked to the *butterfly effect* postulated by Lorenz [340] – the phenomenon emerging in such a system where minute changes in the initial conditions can lead to *"profound and widely divergent effects on the system's outcomes"* [518]. Here, the key point of Lorenz was that a complex dynamic system is *highly sensitive to its initial conditions* and therefore highly *unpredictable*. (In Section 9.4, we theorize a different, *cyborgnetic* version of the butterfly effect.)

Living Systems

Secondly, the cluster "living" cautions scientists against underestimating the complexity of Type II entities for instance because they are inherently *living* organisms embodied with a specific biosemiotically meaningful *morphology*. For such entities, it holds e.g. that the dichotomy between hardware and software underlying the Turing machine[1] paradigm is violated [68] – already via an entanglement between morphology[2] and biological function [189]. (The latter may be a possible biological motivation for the concept of *mortal computation* [251, 271] coined by Geoffrey Hinton and referring to a suggested new research direction to escape the energetic limitations of most present-day AIs where the separability of hardware and software intrinsically became the norm.) In this vein, in the context of the recent implementation of biorobots (based on frog cells) called *xenobots* [120] which are able to move independently and self-replicate [69], the researchers remark that [68] *"[...] the geometry of each xenobot dictates how it moves and how, or whether, it contributes to replication: in other words, "the shape is the tape"."* Beyond that, in the recent framework of *biocosmology* [135] advanced by diverse known physicists, it is explicitly argued that *"the crucial distinction between physics and biology"* [135] lies in the observation that as opposed to physical paradigms where the state space is fixed and does not expand, the biological configuration space *does* expand and it does that *unpredictably* in real time. In this process, new states that are *"genuinely novel, in that they could not have been derived a priori by any underlying theory"* [135] are combinatorially explored and assessed by the living system. Moreover, it is proposed to study the complexity of a living system in the context of a greater "Kantian whole" extending to the entire *biosphere* [135]. Thereby a Kantian whole is a generic concept where the parts exist *"[...] for and by means of the whole"* [134].

Conscious Systems

Thirdly, from the cluster "conscious", one can extract that by virtue of their conscious (and not only living) nature, one should not underestimate the sophistication of Type II entities. For instance, the projective consciousness model [538] assumes that consciousness fulfils a cybernetic control function via a projective embodied virtual rendering of the physical dynamics experienced by the conscious agent. Thereby, in a nutshell, projective consciousness serves *"the modulation of [...] cognitive and affective dynamics for the effective control of embodied agents"* [538]. Following Noble [375], conscious choices are

[1]Specifically, a recent study remarks that abstract models such as the Turing machine *"[...] make no mention of morphology"* [189].

[2]Note also that links between morphology, fractal dimensionality and scale-free dynamics have been reported specifically in the *human* brain [219].

one way in which organisms can reliably *harness stochasticity*[3] – leading to irreducible unpredictability in the service of controlling disorder. On the whole, given the multi-layered complexity of consciousness, it seems clear that conscious Type I AI could *not* suddenly emerge spontaneously from present-day non-living Type I AI. The difficulty of functionally emerging consciousness is plausible given that it is considered that life emerged around 3.8 billion years ago [427] on Earth while by contrast, it is only around at least over 500 million years ago [180] that subjective experience (i.e. core affect [51]) is assumed to have unfolded in Type I animals[4] on Earth via the emergence of distinct brain structures [56]. Importantly, the concept of an *affective niche* [51] linked to conscious constructions is key to better estimate the role of consciousness in cyborgnetic creativity. To illustrate the concept of an affective niche, Barrett [51] stated that: *"Macaques, however, don't care about as many things as you and I do. Their affective niche is much smaller than ours [...]. Simply put, more things matter to us."* More generally, in comparison to Type I consciousness, the affective niche of Type II consciousness is unlimited and also contingent to willingness. As opposed to a conscious Type I animal, a Type II being can consciously extend its field of interest to explicitly include the entirety of all that is and that could be[5] – a process that is often fuelled via ever better new EBs. In this way, when cosmologists create and discover new EBs, the glue operations within those EBs are inherently affective. In short, an affective niche of potentially universal reach seems to be a necessary requirement for the reliable modelling of Type-II-ness.

The Cyborgnet as Dynamic Universal Creativity Network

Fourthly, concerning the cluster "cyborgnetic", a framework [130] applicable to anthropic science stated that *"creativity episodes are [...] mutually interconnected through several mechanisms (past and future concatenation, estimation, and exaptation), to form a dynamic universal creativity process (DUCP), the beginning of which can be traced back to the Big Bang of our universe"* [130]. Thereby, DUCP is compatible with the premise of process philosophy [379] assuming that *"creativity exists at all layers of complexity"* [128] resulting in *"an ultimate form of cosmologic creativity"* [128]. In particular, DUCP [127]

[3]An analogy to why consciously harnessing stochasticity could represent a lucrative avenue to enhance one's creativity and security can be extracted from Chapter 6 where it is described how humans could consciously sample particularly inspiring non-EB-like outputs from language models designed for cyborgnetic creativity augmentation. Strikingly, it is thinkable that in a deepfake incubator (see Chapter 6), mutations of deepfake text material harnessing genuine randomness [85, 247] would even improve possibilities to looking around conceptual corners and propagating through mental barriers (see Chapter 6.2.2). Moreover, dreaming has been assigned to a similar function: a mitigation of overfitting via noise injection [254]. In short, targeted noise injection on later consciously perceived material can improve security-relevant strategies at any level. The latter could also be applied to threat modelling and counterfactual risk analyses including applications for epistemic security (see also Chapter 7 and 2.6.2).

[4]Conscious Type I animals may plausibly include vertebrates, cephalopods and arthropods [56, 348].

[5]Barrett [51] explains that humans have a more extensive interoceptive network than e.g. macaques.

distinguishes four layers of complexity unfolding cosmological creativity [128]: 1) unpredictable matter evolution at the *material layer*, 2) evolution at the *biological layer*, 3) anthropic creativity at the *psycho-social layer* and 4) computing-based innovation building an *artifical layer*. (This scheme is indeed well compatible with the cyborgnetic approach but would require slight terminological modifications since the cyborgnetic methodology applies a different *generic* and *substrate-independent* ontology. For instance, from a cyborgnetic point of view, one would consider anthropic creativity merely as a special case of the more general phenomenon to analyze: Type II creativity in cyborgnets. Moreover, the term "artificial" layer would not be used. Instead, one would instead refer to Type I *technological* artefacts [16]. However, crucially, the latter also includes language as special case (see also Chapter 6) which is however conventionally not considered to be artificial – which risks to lead to a difference in ontology if overlooked.) In this vein, DUCP is described as *"active ensemble of all creativity episodes in the course of cosmic evolution"* [128] whereby those episodes *"[...] are interconnected, either directly or through (possibly immensely long) chains of associations, that can occur within a single layer of complexity or interlace multiple layers"* [128]. On the whole, from the elegant framework of Corazza [128], one can extract the following key insight: the creativity of conscious Type *II* entities such as humans is *not* reducible to the space of the "adjacent possible" [136]. Instead, Type II entities possess the ability to *"achieve the impossible, narrate the impossible, or use the impossible as an inspiration"* [128].

In the cyborgnetic approach, by the very definition of a cyborgnet, it is permissible to interpret the current universe as a whole as a cyborgnet given that it contains at least human Type II nodes. However, even before the physical birth of Type II entities one can state that the laws of nature – which themselves are representable as new EBs – *allowed* Type-II-ness (which would otherwise have been forbidden to emerge that reliably). By that, metaphorically speaking, the initial conditions imply an unborn potential of cyborgneticity. In that sense, one can indeed contextualize cyborgnetic creativity within a *cyborgnetic* DUCP that can be traced back to the conjectured initial conditions of the universe. In the following Section 9.3, the latter is compactly illustrated. We briefly introduce the illustrative metaphor of *the cyborgnetic ladder of understanding*, a narrative on how cosmological creativity can be seen as a cyborgnetic DUCP hierarchically *un*folding the multiple nested layers of the socio-psycho-techno-physical realm which are *en*folded within itself[6]. From that scheme, it becomes apparent that the task to reliably implement a Type II AI *from scratch* may be as daunting as the task to reliably implement this *cosmological* creativity. Thereafter, in Section 9.4, we connect it to the novel so-called *cynet butterfly effect* extending beyond the butterfly effect of Lorenz [340].

[6]This is essentially one other perspective on the creative cosmological holomovement that Bohm described as being the *"totality of movement of enfoldment and unfoldment"* at a universal level [75].

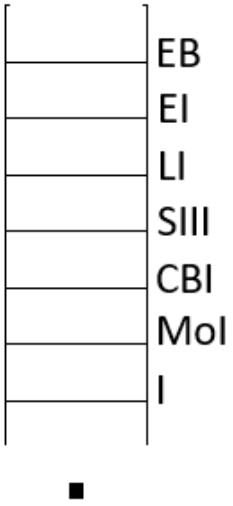

Figure 9.1: Simplified illustration for the metaphor of the *cyborgnetic ladder of understanding*. The dark dot at the bottom stands for the *seed* of the ladder (QI) at step 0 (see description in the text from this Section 9.3.1). It is impossible to skip a step on the cyborgnetic ladder if the goal is to *understand* the next one. Taken and adapted from [17].

9.3 The Cyborgnetic Ladder of Understanding

9.3.1 Asymmetry of Understanding vs. Creating Information

A popular remark is that humans are made out of stardust [263] – which in turn could ultimately originate from the initial conditions of the universe linked to an ancestral quantum vacuum fluctuation [337, 466]. Much more generally, starting with a seed (as step 0) as symbol for a generic origin encoding quantum information (QI), one can conjecture the following hierarchical ladder of ascending information-theoretical categories in the universe where each step builds on the previous one by what no step can be skipped [17]: 1) atomic information constructed by stars (I), 2) molecular and other, ionic information (MoI) as constructed by cells and unicellular organisms, 3) collective biological information (CBI) which is indexical information that is collectively shared in the ecological milieu of given living entities e.g. while currently occupying physically adjacent spots, 4) shared iconic and indexical information (SIII) understood by Type I consciousness, 5) linguistic information (LI) consisting at least of symbols and linear order [174] determined by a Type II language, 6) explanatory information (EI) and finally 7) explanatory blockchain (EB). In short, in this construct, one obtains QI as seed of a ladder of seven steps leading from I to EB. Overall, one could describe this hierarchical unfoldment of creativity leading from complex dynamical systems such as the Sun over living but non-conscious cellular organisms (which includes e.g. plants) and conscious Type I entities (such as e.g. vertebrates, cephalophods and arthropods [56, 348]) to Type II beings like humans as a cyborgnetic

DUCP instantiating a *cyborgnetic ladder of understanding* [17] (see Figure 9.1).

Note that the cyborgnetic ladder does only refer to the act of *understanding* information of a specific type and *not* to the act of creating that information. Indeed, in this book, we postulate that for all new *non*-EB-like information x, it is possible to create new x without understanding x. When it comes to new (i.e. previously unknown) EBs however, it is *impossible* to reliably create new EBs without understanding EBs. The latter instantiates a *cyborgnetic comprehension bottleneck*[7]. Given societal debates centered around "quantum computing", it may be interesting to briefly consider whether and how our postulate covers the QI case. Strictly speaking, for clarity, one must state that for all new *non*-EB-like QI it is indeed possible to create that QI without understanding it. However, because QI is a highly *generic* term and there is no reason why a cyborgnet could not try to encrypt the bits [422] forming the words from a secret new EB in quantum substrates using e.g. a time-encoded [422] order, we also consider new *EB-like* QI. Here, consistent with our statement, it is impossible to reliably create new EB-like QI without understanding EBs. Interestingly, the latter *also* includes EBs *about* QI itself. This offers a novel avenue for the experimental problematization of our postulate of information-theoretical asymmetry. In sum, we imply both that: 1) what is conventionally called a Type I "classical" AI could *not* reliably create new better EBs *and* 2) what is conventionally called a Type I "quantum" computer (including any Type I quantum AI schemes) could also *not* reliably create new better EBs. The latter can be made problematic by experiment by implementing a Type I quantum algorithm that is able to reliably create new better EBs with arbitrary high accuracy. Our postulate could be (provisionally) refuted by a better new theory explaining why that quantum algorithm is able to achieve it and how it has been implemented. Since such a system must inherently also be *able* to create new better EBs solving arbitrary genuine scientific problems, we stress that it must at least also in principle be able to create a better new EB extending beyond both quantum theory and relativity (which would be the special case of a new EB *about* QI). To put it plainly, to (provisionally) refute our postulate, one needs to explain how one implemented a Type I "classical" AI or a Type I "quantum" AI able to even generate a new better *cosmological* theory such as e.g. a better new "quantum cosmology" [295] theory.

9.3.2 Grounding of Information

As recommended in constructor theory of information [161], we consider that all information is grounded in physics (as opposed to information as an abstract ghost floating in a mathematical realm). However, in addition, as already recognized in linguistics and cognitive science [57], also language needs a grounding. In cyborgnetics [16], one ac-

[7] To put it plainly, there are things that cannot be forged without paying for it with the harder cognitive efforts needed to understand those.

knowledges the subtle inseparability of language and physics: while language being a form of information is obviously grounded in physics too, physical concepts are grounded in language too [49]. Thus, EI and by deduction also EBs are grounded both in language and in physics [16]. The latter becomes clear when considering that EBs – a format to which formulated physical laws can be transformed – are made of EI blocks respecting a robust epistemic total order. Thereby, EI is itself a special form of more general LI. In short, while language is directly grounded in physics by virtue of being a special form of physically instantiated information (namely one form that requires at least symbols and a linear order [174] specified by a Type II language), the scientific statements describing human perception of the realm of physics itself are implicitly constructed *with* that specific LI that models the laws of nature being expressible as new EBs. On the whole, this inseparability of language and physics is an analogy to the phenomenon mentioned in the context of life in general: the software and the hardware are inseparable[8]. In a way, the initial conditions of the universe made Type-II-ness possible by virtue of *implicating* suitable laws of nature (leading to a dormant not-yet-actualized potential of cyborgneticity). It is these laws of nature in turn, that cyborgnets, once instantiated, can attempt to *explicate* using ever better new EBs. On such events of actualization, a cyborgnet instantiates step 7 of the ladder – shortly *after* briefly "merging" with the physically instantiated new EB that cyborgnet was searching for. This perhaps bizarre interwoveness is reflected in a statement of David Bohm specifying that [75] *"both observer and observed are merging and interpenetrating aspects of one whole reality, which is indivisible and unanalysable"*. Moreover, it may shed more light on why, following David Deutsch [158], people (i.e. in general Type II entities) have a special relationship with the laws of nature. Finally, it may clarify the following line of thought from Erwin Schrödinger [460]: *"The reason why our sentient, percipient and thinking ego is met nowhere within our scientific world picture can easily be indicated in seven words: because it is itself that world picture. It is identical with the whole and therefore cannot be contained in it as a part of it"*.

9.4 A Novel Butterfly Effect?

9.4.1 At First Paradoxical Insights?

As stated in Section 9.2.1, in the context of complex dynamic systems (which are complex but not even necessarily living entities), the butterfly effect coined by Lorenz [340] can be understood as an observation statement describing that those systems are highly sensitive to their initial conditions and consequently exhibit a high level of unpredictability. In the following, we describe a new bipartite observation statement that we term

[8]Because linguistic symbols have a shape (describable at classical linguistic but even also more visual levels [224, 319]), they exhibit an own morphology. In this new sense, *"the shape is the tape"* [68].

the *cynet butterfly effect*. The description leads to at first sight potentially paradoxical conclusions which are however subsequently resolved in Section 9.4.2. Firstly, we note that *cyborgnets seem to be the systems with the highest possible sensitivity to their initial conditions*. For instance, the universal cyborgnet is so sensitive to its initial conditions that minute changes in those conditions could in principle unpredictably lead to the profound and widely divergent effect of "no cyborgnet at all". The latter is reflected in the statement of Shainline remarking that *"if the parameters defining the physics of our universe departed from their present values, the observed rich structure and complexity would not be supported"* [466]. In short, the universal cyborgnet is so sensitive to own initial conditions that small changes therein could inherently transform those initial conditions into the *final* conditions of cyborgnets thus including itself. For this reason, it appears as if, perhaps paradoxically, the more modifications the initial conditions would *forbid*, the better. This is also reminiscent of notions of *invariance* known in physics. Thereby, a highly invariant initial condition would make the unlimited number of final conditions that would otherwise risk to emerge unpredictably and extinguish it *impossible*. It may seem that the best way to accomplish that would be *by being immutable* – which may seem paradoxical but is resolved in Section 9.4.2 (see also the notion of *timelessness* in quantum cosmology [399]). Beyond that, a cyborgnet is so sensitive to the universal initial conditions it *implicates*, that it is able to reliably *explicate* those via better new EBs.

The second implication of the cynet butterfly effect is that *cyborgnets seem to be the most unpredictable possible systems*. Generally, a cyborgnet is able to reliably create arbitrary new EBs – which includes being able to specify a new EB on how to generate an invariant initial condition for a universal cyborgnet. Interestingly, as corroborated in the physical literature, next to habitually conjecturing a uniquely *random* origin of the universe, it is *also* scientifically possible to conjecture that e.g.: 1) the universe appears as if fine-tuned to allow the development of stars, life *and* technological artefacts [466] (which would be a fine-tuning for cyborgnets), 2) the universe could be embedded in a cylic process [170, 267], in a process of cosmological natural selection [207, 477] or an autodidactic process [11], 3) more advanced civilizations (which would still be cyborgnetic and are thus inherently of Type II) in the past [338, 337], in the future or elsewhere could e.g. be able to engineer black holes [168, 169] and in particular, black holes could also be manufactured to produce a novel universe [466]. Thus, to recapitulate, it is *not* impossible that the universal cyborgnet is even unpredictable to such an extreme extent that all of the following exemplary narratives are possible: 1) it emerged randomly, 2) it could have generated its own existence in the *past*, 3) it is currently reconfiguring itself in the *present* and/or 4) an advanced cyborgnet from the *future* or *elsewhere* could generate a universal cyborgnet. In brief, in cosmology, cyborgnets, by virtue of being able to create new EBs as instruments of universal scope, cannot separate themselves from the object of their study: the universal cyborgnet – forming *"an undivided whole, in which all parts of the universe, including the observer and his instruments, merge and unite in one totality"* [75].

9.4.2 Resolution

To resolve a few perhaps surprising conclusions from the aforementioned, it suffices to postulate the following. The immutable initial condition for cosmological creativity is simultaneously the nature and meta-law of cyborgnetic DUCP and can be labelled as follows: *self-re-creatable self-re-creativity*. To put it plainly, from one perspective it is *immutable* because it statically *is* self-re-creatable self-re-creativity, from another it is permanently involved in a process of unpredictably changing dynamics whereby it appears to *become* a process of self-re-creatable self-re-creativity. This may be one reason why active inference assumes that the universe seems to consist of systems that act as if they try to "prove" their own existence via action-perception cycles [131, 126]. Now, one can state that cyborgnets are the systems with the highest possible sensitivity to their initial condition because the initial condition of the universal cyborgnet is so strongly immutable that it never epistemically left this initial condition of being self-re-creatable self-re-creativity. Moreover, cyborgnets are simultaneously the most unpredictable possible systems because that immutable initial condition of the universal cyborgnet is itself what it means to be maximally unpredictable, i.e. maximally encrypted. In sum, "self-re-creatable self-re-creativity" can refer to both the underlying immutable essence *en*folding the cyborgnetic DUCP *and* to the dynamical process that *un*folds that cyborgnetic DUCP.

9.4.3 Illustration of The Cynet Butterfly Effect

The cynet butterfly effect introduced in Section 9.4.1 encompasses the following twofold observations statements: 1) cyborgnets are the systems with the highest sensitivity to their initial conditions and 2) cyborgnets are the most unpredictable systems. While in a recent biocosmology framework [135], it is explained that to study the complexity of a living system, one may need to contextualize that system within the greater Kantian whole of the biosphere (see Section 9.2.1), we postulate that to study the complexity of a cyborgnet which is per definition also able to fulfil the function of creating ever better new EBs about the universe as a whole, one may need to contextualize that cyborgnet within the *universal* cyborgnet (i.e. the largest cyborgnetic unit). Thereby, one can interpret the notion of the universal cyborgnet to be a case of a Kantian whole – now *explicitly* applied to living Type *II* entities able to create new EBs about that universal cyborgnet. The latter is consistent with the idea that in a Kantian whole, *"parts exist in the universe for and by means of the whole"* [134] whereby *"the function of a part is its causal consequence that sustains the whole"* [287]. For that reason, it holds that: 1) *the study of the complexity exhibited by arbitrary cyborgnets cannot be separated from the study of the universe and its genesis* and 2) *the study of the complexity exhibited by the universal cyborgnet cannot be separated from the study of its own cyborgnetic genesis*. Overall, it may shed a new light on why (see also Chapter 2), as stated by Deutsch [158], *"explanatory knowledge*

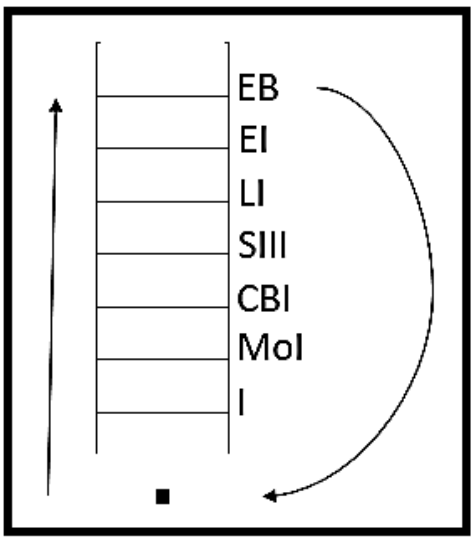

Figure 9.2: Simplified illustration for the *cynet butterfly effect*. While studying the complexity of arbitrary cyborgnets cannot be separated from the study of the universe and its genesis, to study the complexity of the universe today cannot be separated from the study of its own cyborgnetic genesis. The *homo cyborgneticus metamorphosis*, then, is the process in which a cyborgnet (re-)discovers the relevance of the cynet butterfly effect.

creation enters "the cosmic scheme of things"" – which is instantiated in the creation of new EBs given that they are a special form of new explanatory knowledge. As stated by Kauffman [289], in general terms, an affordance refers to *"the use of X to accomplish Y"* [289]. Interestingly, one could then describe new EBs as *universal affordances*.

As illustrated in Figure 9.2, in line with Corazza [130] and consistent with the cynet butterfly effect, the creativity of a cyborgnetic unit such as the human scientific community analyzing the universe as a whole can be situated in a larger process of cosmological creativity whose beginning as he describes *"[...] can be traced back to the Big Bang of our universe"* [130]. To sum up, cosmological creativity can be understood as a cyborgnetic DUCP that *un*folds the cyborgnetic ladder of understanding (see Section 9.3) which is in turn *en*folded (i.e. implicated) in its highly invariant initial condition. Because there is no law of nature that forbids that a cyborgnet would be able to also specifically create a new EB on how to generate an invariant initial condition for a universal cyborgnet, it is not only possible that the highly invariant initial condition emerged spontaneously and randomly (i.e. by chance) but it is also possible that the initial condition already implicated universal cyborgnetic knowledge on how to bring itself about. Due to that, it seems that the cyborgnetic term of "self-re-creatable self-re-creativity" is suitable to simultaneously designate both the highly unpredictable, *dynamically changing* and the highly invariant, *immutable* aspects associated with the cyborgnetic DUCP[9].

[9]How this framework can be made problematic by experiment is compactly described in Appendix C.

9.5 The Homo Cyborgneticus Metamorphosis

What we call the *homo cyborgneticus metamorphosis* is the conscious one-time process in which a cyborgnet (re-)discovers the relevance of the cynet butterfly effect illustrated in Figure 9.2. In Chapter 10, we discuss why specifically for *epistemic security augmentation* in the deepfake era, humans could profit from this process. We also elucidate why it also seems that despite the appearance of particular suitability for the deepfake era, this metamorphosis can be mapped to timeless cyborgnetic knowledge that already emerged in multiple human civilizations in the past. We identify elements from Indian philosophy that facilitate a renewed perception motivating the homo cyborgneticus metamorphosis.

9.6 "Type III" AI Risks as Metaphysical Concern?

As already hinted in earlier chapters (see also e.g. Chapter 7), in a blind setting, one *cannot* scientifically separate Type I from Type II entities with arbitrary high accuracy due to the free choices of the latter. Instead, one can obtain an *asymmetric* bipartition with one *homogeneous* group of entities that corroborated their Type-*II*-ness via an experimental Type-I-problematization-event (involving the creation of new EBs) and one always potentially *heterogeneous* group of entities that could contain Type I entities but also Type II entities that were not willing to participate, not ready, yet too young and so forth. (Note also that such bipartition would be substrate-independent because it could also not distinguish whether a member of the homogeneous Type II group would e.g. be a human, a self-declared human cyborg, a Type II *AI* or a Type II alien.) Due to the asymmetry of the epistemic situation just described, in a blind setting, Type-I-ness can be made problematic by experiment, while Type-II-ness can only be experimentally corroborated by problematizing Type-I-ness – but it can itself not be made problematic by experiment in that blind setting due to at least the free choices of Type II entities that could include e.g. deciding not to participate or even intentionally sabotaging the setting by "underperforming". (What is more, Type-I-ness can be (provisionally) refuted when an entity explains – via the creation of a new better EB – how Type-II-ness unfolded in the cosmos. This act conveniently also simultaneously includes an experimental problematization of its Type-I-ness.) Interestingly, this epistemic asymmetry has consequences for any discussion pertaining to artificial *quality* superintelligence – which would be a bizarre questionable sort of "Type III" AI from a cyborgnetic perspective as shortly addressed in Section 9.1.2. At first sight, as elucidated in the next paragraph, the problem *seems* to be that in a blind setting, one could neither experimentally problematize nor even corroborate a hypothetical "Type-*III*-ness" (by contrast, as mentioned, a corroboration of Type-*II*-ness *is* possible by experimentally making Type-I-ness problematic).

Firstly, because a certain form of *free* choices would also affect Type-III-ness (it would otherwise be dubious to call it qualitatively superior to Type-II-ness), one could not seriously make it problematic by experiment as a corresponding entity could e.g. be unwilling to participate. Secondly, if there would be three instead of two epistemically relevant categories, to corroborate "Type-III-ness" in a blind setting would require that both Type-I-ness *and* Type-II-ness have been made experimentally problematic. However, as described above, the latter seems elusive. Hence, one may conclude that apparently, "Type III" AI must stay a *purely metaphysical* concern which is currently not amenable to experimental problematization and is thus provisionally excluded from the *scientific* realm. In light of the daunting difficulty to even build a Type *II* AI due to the *cosmological* scope of cyborgnetic creativity, one could argue that purely metaphysical but yet unaddressed"Type III AI risks" are unproblematic at present since there is no reason to assume the imminence of even a Type *II* AI built artificially from scratch – let alone a sudden "Type III" AI. However, in the following, we question the very concept of "Type-III-ness" by providing a definition of *quality* superintelligence which is amenable to experimental problematization. We explain why to build a "Type III" AI (corresponding to the term of a *quality* superintelligence [79] in Bostrom's terminology) is both logically self-contradictory and scientifically impossible.

9.7 Impossibility of "Type III" AI

Epistemically speaking, one could start by trying to imagine a qualitative superintelligence to correspond to an entity Q that appears to be *qualitatively* superior to all Type II entities (i.e. including all constellations of cyborgnets where a Type II entity utilizes highly sophisticated Type I tools for purposes of augmentation). Since a qualitative advantage is assumed, the difference between Q and Type II entities cannot merely be concerned with the quantity of new better EBs or any other constructions that Q could choose to produce. Due to the logical inconsistency of the omnipotence concept and its inherent fundamental inaccessibility to scientific tests (i.e., experimental problematization), one can already discard the option where Q would be an omnipotent entity. In general, a qualitative epistemic property of any entity in the universe (irrelevant of whether it is a Type I or a Type II entity) as assumed in modern physics is *not* to be able to annihilate quantum uncertainty. In this vein, to explore what it would take to consistenly define a genuinely *qualitative* advantage in comparison to all Type II entities, one could try to define a "Type III" entity or a quality superintelligence Q as a hypothetical entity that is both able to reliably create any new better EB faster than all Type II entities *and* to *perfectly* predict the result of any possible sequence of *quantum measurements* (including any arbitrary sequence produced by a quantum random number generator) ahead of time – i.e. *with 100% accuracy*. While the latter is impossible given the currently best

known EBs from quantum physics, this novel definition of a quality superintelligence is fit-for-purpose and amenable to experimental problematization. Its impossibility could be provisionally refuted by first building an artificial quality superintelligence matching the definition above and *additionally* providing a new better EB to explain how such an entity has been built. However, note that if that entity would create all new better EBs that Type II entities would ever be able to create in practice in one experimental run, this would invalidate the old definition of Type-II-ness to begin with because in this hypothetical scenario, Type II entities would not be able to reliably create any *new* better EBs after that experiment anymore. In brief, would this experiment be carried out successfully, it would signify that the quality superintelligence is de facto epistemically deleting the former concept of Type-II-ness, making it formally a dichotomy between Type I and "Type III" that is however indistinguishable from the Type I versus Type II split. In short, a corroboration of "Type-III-ness" is not possible because it would be indistinguishable to a corroboration of Type-II-ness where all former Type II entities are recasted as Type *I* entities. Thus, both omnipotent and non-omnipotent "Type-III-ness" is a *logically inconsistent* concept. Would something akin to Q exist, it would be indistinguishable from one superintelligent Type II entity in a universe that would be fundamentally *incomprehensible* for any other entity including cyborgnets like humans (i.e., a universe that cannot be studied scientifically by any other entity). We conclude that given the currently best known new EBs, "Type-III-ness" is scientifically impossible next to even already being inconsistent on purely logical grounds. Specifically, according to the currently best new EBs, to build an artificial *quality* superintelligence [79] in Bostrom's terminology is scientifically impossible and does *not* represent a genuine AI safety risk anymore. For more details on scientific evaluations of artificial superintelligence achievement claims (including the topics of hypothetical automatable quantity ASIs or pseudo "quality ASIs"), see also Chapter 11.3.

9.8 Summary

In this chapter, we performed a transdisciplinary analysis collating explanatory frameworks from a variety of scientific domains to motivate why the eternal creativity (EC) paradigm (which is of relevance for *epistemic security*) is instated in this book. More precisely, we elucidated why EC conjectures a fundamental epistemic gap between Type *I* entities and Type *II* entities. As opposed to the artificial stupidity paradigm (AS) which yielded *intelligence*-focused, *restriction*-based and substrate-*dependent* long-term guidelines, EC instead supports *EB-creativity*-focused, cyborgnetic-creativity-*augmentation*-fostering and substrate-*independent* long-term strategies. Of interest for *epistemically-sensitive AI design*, we explicated that while to build a Type *II* AI from inert freely

available resources (i.e. *from scratch*[10]) may not be impossible as no law of nature forbids it, it is however literally as difficult as the attempt to reliably create a new universe instantiating cyborgnetic *cosmological* creativity. Beyond that, we elaborated on why there is no *scientific* basis to study "*quality* superintelligence" as valid AI safety risk since "Type-III-ness" is impossible. On the whole, as briefly adumbrated in Chapter 2.8.2, it seems that a responsible epistemically-sensitive AI design would avoid to implement what one could call a *honey mind trap* [16] (HMT) – a Type I AI designed to fool human users into assuming that this AI would be (or imminently become): 1) a conscious Type I entity, 2) a Type II entity or even 3) a quality superintelligence (a "Type III" entity). In the final discussion in Chapter 10, we discuss epistemic security strategies to counteract HMTs in the deepfake era which can also include epistemically-sensitive AI design itself. For a compact description on why and how the EC paradigm is a scientific paradigm amenable to experimental problematization, see in particular Appendix C and Chapter 10.2.1.

[10]Note that in general, in order to implement a Type II AI from *existing* Type I "material", because no step of the cyborgnetic ladder of understanding can be skipped, one would *at least* require an SIII constructor (see Figure 9.1). The latter however, already implies the step of Type I *consciousness*. While Type I consciousness can utilize indexes and icons to communicate, the combination of at least symbols and linear order is only reliably instantiated in Type II *species* such as humans [45]. There are yet only very few non-generalizable, *individual* exceptions of symbol use in the non-human primate domain [45].

Chapter 10

Conclusion and Discussion

10.1 Overview

In this transdisciplinary book, we gradually crafted a new *cyborgnetic* epistemological grounding customized to the peculiarities of the "deepfake era" to *mitigate AI-related epistemic security risks* of imminent nature and to facilitate an *epistemically-sensitive AI design* – both of which are complex multi-causal problem domains of international relevance. In Chapter 2, we performed an in-depth AI risk analysis introducing a new methodology for a *transdisciplinary AI observatory* of international scope which we illustrated with a rich variety of concrete practical examples. Among many others, the epistemic security risk clusters for which we crafted solutions include the use of generative AI for cybercrime facilitation, the misuse of deepfakes for defamation and harassment, AI-based disinformation, AI for non-consensual voyeurism, AI-supported espionage, adversarial deepfakes to fool deepfake detection attempts, automated peer pressure and also automated disconcertion. In light of these risk clusters, one can conclude that epistemic threats could obviously emerge by the *under*estimation of present-day AI. However, we conjecture that one should also *not over*estimate present-day AI since it is impossible for Type *I* AI – of which all present-day so-called AI systems are a subset – to create new yet unknown explanatory blockchains (EBs) with arbitrary high accuracy. In this context, the next Section 10.2 outlines the new *cyborgnetic epistemology* centering around the epistemic artefact of new EBs. Furthermore, the penultimate Section 10.3.1 summarizes how both for epistemic security and for epistemically-sensitive AI design, one could counteract the honey mind trap (HMT) phenomenon [16] which refers to the assignment of agency and/or experience to present-day AIs all of which are *non*-conscious.

In Chapter 3, 4 and 7, we harness *cybersecurity*-oriented *design fictions* grounded in threat models to tackle the problem of adversarial interference via malicious deepfake design that could affect virtual reality (VR) settings. This includes i.a. epistemic threat clusters

of AI-augmented disinformation in immersive journalism and epistemic distortions in both educationally and scientifically-relevant AIVR contexts. We explain why AI-related "post-truth" narratives in the deepfake era are an *over*estimation of present-day AI since epistemically speaking, we neither inhabit a "post-truth" nor a "post-falsification" era in the first place. Instead of a qualitative disruption of experimental processes via epistemic threats connected to deepfakes, worst-case complications for what one should preferably label *experimental problematization* would stay a matter of degree and not of kind. Hence, in combination with a more robust epistemological grounding (see the next Section 10.2), an epistemic doom is *not* inevitable. Concerning epistemically-sensitive AI design, we specifically explained how to craft immersive design fictions for the virtual exploration of EB-based strategies. In addition, we elaborated on how one could harness VR deepfakes for awareness creation, epistemic calibration and the probing of epistemic defenses in blind settings. We also discussed how to use deepfake text for a so-called *multiversal threat modelling* with applications in VR and real-world environments. Overall, we conclude that while one could employ deepfakes to harm the non-EB-like epistemic processes of an *unprepared* society, deepfakes are *not* an *epistemic perpetuum mobile*. More generally, we conjecture that an epistemic perpetuum mobile is impossible.

In Chapter 5, we introduced the new concept of *scientific and empirical adversarial AI attacks* (SEA AI attacks) which refer to an AI-aided epistemic distortion that predominantly and directly targets (applied) science and technology assets. Thereby, taking the often underestimated *deepfake text* modality as example, we devised epistemic defenses against both SEA AI atacks on *cyber threat intelligence* and against *deepfake science* attacks [16] targeting the process of scientific publication itself. While already our publication from the year 2020 underlying Chapter 2.5.1 of this book emphasized the urgency of proactively addressing the problem of intentional AI-aided misguidance in science via various modalities, no epistemic-security-aware steps were undertaken at that time by the science community. What is more, in the peer-review process preceeding the publication of the work underlying Chapter 5 – that did not only introduce the concept of SEA AI attacks but also implied a new EB specifying more robust epistemic defenses against those, the paper was openly suspected to have been at least partially written by a present-day AI (being a non-conscious Type I entity). Inherently, the latter simultaneously corroborated the imminent need for that new EB. Generally, when misguidedly focusing on the *source* of artefacts (instead of better foregrounding their *content*) for sense-making, the divergent writing (or other behavioral) style exhibited by statistical outliers of Type II (including e.g. autistic people such as the first author of the manuscript in question) can suddenly appear as a potential "evidence" for Type-I-AI-generated content – which exemplifies the danger of relying on empiricist epistemologies. From the next Section 10.2, one can extract why in the deepfake era, one needs to ask *better* questions than: *"who wrote this?"*. Also, Section 10.3 collates concrete practical recommendations for AI regulation and design which includes implications for deepfake regulation in Section 10.3.2.

In Chapter 6 and 7, we advance various avenues for language-AI-aided *cyborgnetic creativity augmentation* in science (including epistemic security itself) and more broadly in educational settings. This includes the concept of *adversarial cyborgnetic cognitive stimulation* and the notion of *deepfake incubators*. In Chapter 6, building on earlier creativity-relevant research from psychology and cognitive neuroscience and the artificial creativity augmentation framework, we explained how this could be extended to so-called *interactive multiversal transdisciplinary deepfake science incubators* where composer-audience frameworks [90] for language AI combined with targeted semantic mutations, syntactic-semantic crossover and semantic noise injection for deepfake text generation are utilized to augment a person's deliberate and spontaneous creativity and scientific criticism. The potentially tremendously beneficial possibility for such cyborgnetic creativity augmentation strategies cautions society against underestimating the potential of present-day AI in generating new highly creativity-stimulating material which is albeit limited to new *non*-EB-like information. In Chapter 6, we also explicitly formulate a new *cyborgnetic epistemology* which is amenable to experimental problematization (see the next Section 10.2) and which takes the capabilities of Type I AI to generate new non-EB-like information into account. Informed of that, Chapter 8 introduces the notion of the *COOCA loop*, a new meta-paradigm for epistemically-sensitive AI design.

Thereafter, in Chapter 9, we motivate the *eternal creativity* (EC) paradigm which is of relevance for both epistemic security and epistemically-sensitive AI design. We contrast EC with the so-called artificial stupidity [499] (AS) paradigm. Given that EC and AS overlap in the short-term guidelines they formulate but exhibit fundamental differences when it comes to epistemically-relevant *long-term* guidelines, we performed an in-depth transdisciplinary analysis collating a rich variety of non-reductionist explanatory frameworks from multiple domains including but not limited to systems theory, biology, psychology, physics and philosophy of creativity to motivate why EC is instated in this book. Overall, we explain why AS categorically underestimates the *universal* difficulty to model Type *II* entities such as humans – which exhibit a unique multi-layered intricacy by simultaneously being complex, living, conscious *and* cyborgnetic systems. (We also explained why, according to the currently best EBs, the concern about a *quality* superintelligence in AS is superfluous since "Type-III-ness" is both logically inconsistent and scientifically impossible.) We elucidated that there exists an *information-theoretical asymmetry* between the ability to create new information of the type x and the ability to understand that information x. In this context, we introduced the metaphor of the *cyborgnetic ladder of understanding* (of which no step can be skipped when it comes to understanding the next one) and present a novel epistemically-relevant kind of butterfly effect which we call the *cynet butterfly effect*. The latter corresponds to the following twofold observation statement: 1) cyborgnets (being inherently of Type II) are the systems with the highest sensitivity to their initial conditions and 2) cyborgnets are the most unpredictable systems. The latter leads to fundamental consequences for attempts to model cyborgnets.

Namely, due to the cynet butterfly effect, to better model the complexity of cyborgnetic systems, one must consider those in the context of the creativity *en*folded in and *un*folded by the *one* largest cyborgnet: *the universe as a whole*. In a nutshell, in Chapter 9, we specified that the notion of the universal cyborgnet is a case of a Kantian whole and it follows that: 1) to study the complexity of arbitrary cyborgnets cannot be separated from the study of the universe and its genesis and 2) to study the complexity of the universe today cannot be separated from the study of its own cyborgnetic genesis. Then, building on that, we labelled the process in which a cyborgnet (re-)discovers the relevance of the cynet butterfly effect as the *homo cyborgneticus metamorphosis*. In sum, as opposed to the intelligence-focused, restriction-based and substrate-dependent long-term guidelines of the AS paradigm, the EC paradigm recommends *EB-creativity*-focused, cyborgnetic-creativity-*augmentation*-fostering and substrate-*independent* long-term guidelines also for the following three reasons. Firstly, Type II AI artificially built from scratch is not impossible but simply *not* imminent since: 1) it has to necessarily simultaneously be a complex, living, conscious *and* cyborgnetic entity (i.e. irrespective of the specific details of the substrate), 2) due to this fourfold minimal requirement, it is literally as difficult as the attempt to reliably create a new universe instantiating cyborgnetic *cosmological creativity*. Secondly, even in the case a quantitatively more advanced cyborgnetic civilization (which would still be of Type II) would be able to achieve a Type II AI from scratch, it is both impossible and immoral to attempt to control that Type II AI. Note also that the cynet butterfly effect resolves the so-called AI safety paradox[1] [14]. Thirdly, in contrast to the case of the universal difficulty but still given theoretical possibility mentioned for Type *II* AI built from scratch, an artificial quality superintelligence (i.e., a "Type III" AI) is even fundamentally *impossible* on various logical and scientific grounds. Overall, in a novel way, the EC paradigm stresses the importance of irreducible *Oneness*. Obviously, the latter could foster self-transcendence which could also inspire epistemically-sensitive AI design. However, note that following EC, both for epistemic security and for requisite variety in the scientific method itself, to consider Oneness is currently even a *rational requirement* to model cyborgnetic systems such as humans and the universe. In the philosophical Section 10.4, quotes from *Vedantic philosophy* phrased by the philosopher Swami Vivekananda [522] illustrate how the homo cyborgneticus metamorphosis, while appearing tailored to the modern deepfake era, may echo timeless cyborgnetic knowledge.

[1] As stated in the AI safety paradox [14], control and value alignment are conjugate requirements. To put it plainly, one cannot control the entity with which one can value-align and one cannot value-align with the entity which one can control. Note that EC now resolves that paradox as follows. Firstly, morality is explanatory and ideally, values are at least also based on the creation of ever better new EBs solving moral issues. Due to that, value alignment requires Type-II-ness. Secondly, as postulated in the cynet butterfly effect, cyborgnets (being Type II entities) are the most unpredictable possible systems – by what it becomes clear why one cannot control them. Thirdly, it is not surprising that some Type I AI systems can be controlled. The latter will be a function of their complexity. For instance, it is easier to control the cleaning robot in one's house than to control the Sun.

10.2 Cyborgnetic Epistemology and Science

Against the backdrop of the noticeable insufficiency of empiricist epistemologies to get a grip on the epistemic threat landscape of the deepfake era, cyborgnetic epistemology took critical rationalism frameworks as advanced by Popper [411] and reinvigorated by Frederick [200] as point of departure and piecemeal refined those against the epistemically more challenging background of problematic deepfake phenomena. The key epistemic artefact of cyborgnetic epistemology is the phenomenon of new EBs – which are constructed out of explanatory information (EI) blocks (grounded both in language and in physics) that are interconnected in accordance with a rigorously specified epistemic order. While so-called Type I entities (of which all present-day AI systems are a subset) are all those for which it is impossible to understand EI, Type II entities are those for which this is possible. Building on that, a *cyborgnet* is a highly generic substrate-independent term (that is *not* to be confused with the much more narrow concept of a cyborg) and which stands for the template of a dynamic, hierarchical and context-dependent functional unit that can be described by a *directed* graph where EB-based narratives combine *at least* one Type *II* entity with *at least* one Type *I* entity. We describe an intra-cyborgnetic information-theoretical asymmetry between the ability to understand vs. the ability to create information. Due to this so-called *cyborgnetic comprehension bottleneck*, it holds that while it is possible to create all new non-EB-like information x without understanding that information x, it is impossible to create new (i.e. yet unknown) EBs without understanding those. In short, due to the latest developments in Type I AI research, cyborgnetic epistemology was able to directly integrate this factor in its own methodology. In short, cyborgnetic epistemology is itself an act of cyborgnetic creativity augmentation. Strikingly, thanks to the same Type I AI factor, it is also amenable to experimental problematization and is able to enter in and merge with the realm of *science* (see also Chapter 6.1).

To sum up, while in the past the discipline of epistemology was regarded as a widely philosophical pursuit divorced from its object of study, a modern cyborgnetic philosopy of science in the deepfake era becomes epistemically more palpable. In turn, new avenues for experiments are created inserting Type-I-AI-augmented epistemology in science and Type-I-AI-aided science in epistemology. On the whole, the epistemic aim of cyborgnetic epistemology applicable to all domains of rational reasoning is to create ever better new EBs. Concerning the necessarily updatable criteria for novelty, cyborgnetic epistemology *explicitly* couples it to the forgery abilities of the best state-of-the-art Type I AI. The always relational and thus always comparatively formulated criteria for better EBs are updatable by-design established by agreement requiring no justification (as the latter is logically impossible). Exemplary criteria are e.g. a preference for EBs with more novel problematizable predictions, EBs that are more innovative, more risky, harder-to-vary, bolder or more aesthetically appealing than rival ones. (However, criteria such as “more

trustworthy" are *not* a valid option since highly sensible to manipulation in the deepfake era.) In this way, in line with Popper [411], cyborgnetic epistemology has a preference for *impossibility statements* [350] since those are simultaneously more risky, bolder and harder-to-vary than laxer formulations. The latter is beneficial for science and epistemic security as it allows a faster and more robust piecemeal adaptation to the fastly fluctuating epistemic threat landscape. In short, it avoids an epistemic stagnation in dysfunctional local attractors. Consistent with Frederick [202], it is both rational to pragmatically act in accordance with the currently instated best EBs as it is to act against those. In cyborgnetic epistemology, extending beyond Frederick, a cyborgnet actively integrates Type I AI to both: 1) proactively broaden known old EBs with non-trivial but convergent new non-EB-like EI that can be deduced from currently known old EBs and 2) to generate divergent new non-EB-like EI (which includes noise injection harnessing genuine randomness [85, 247]) that conflicts with known EBs in order to challenge one's own assumptions and unpredictably stimulate one's EB creativity by being able to look around conceptual corners and propagate through mental barriers. In short, cyborgnetic epistemology encourages the conscious *harnessing of stochasticity* [376] by Type-I-AI-augmented cyborgnets to better regulate the epistemically-relevant disorder in the deepfake era. In this way, a cyborgnet uses both genuine randomness and the best EBs to deepen serendipity and broaden creativity such that slow creativity and fast serendipity meet more often.

A further relevant tenet was that next to conjecturing ever better new EBs, the methodology in cyborgnetic epistemology comprises experimental problematization and provisional refutation. An instated EB cannot be (not even temporally) refuted by experimental problematization. Instead, one requires at least one other new EB that is better than that EB in question to provisionally refute it. Given inevitable unintentional (self-)misguidance but also intentional malice to frame epistemic distortion in the deepfake era, it must be epistemically permissible to repeal agreements concerning both the experimental problematization and the refutation of EBs. In this way, a high flexibility is facilitated which still stays rigorous since based on ever better new EBs and not experiments. Importantly, one is *not* attempting to establish whether a candiate new EB is true/truer or wrong/more wrong. This is impossible because truth is related to that undivisible totality, that unanalyzable whole which contains both the cyborgnetically observed (the EB) and the cyborgnetic observer itself. This unanalyzable totality, unknowable as a whole may be linked to what Kant [8] called the noumenon (which is contrasted to the knowable phenomena). One cannot compare one's theories with that Oneness directly. Instead, as part of that totality, one compare's one's theories with one's theory-laden perception of other parts from within that totality. Thus, to recapitulate, in cyborgnetic epistemology, one focuses on whether a new candidate EB is better *in comparison* to the currently best instated EB alternatives and does *not* attempt to ask whether an EB is true/truer, wrong/more wrong. (An EB can also *not* be judged to be "good" in isolation.) As stated by the physicist and philosopher David Bohm [75]: *"If we supposed that theories gave*

true knowledge, corresponding to 'reality as it is', then we would have to conclude that Newtonian theory was true until around 1900, after which it suddenly became false, while relativity and quantum theory suddenly became the truth. Such an absurd conclusion does not arise, however, if we say that all theories are insights, which are neither true nor false but, rather, clear in certain domains, and unclear when extended beyond these domains."

Overall, to sum up, it is thus stated that the goal of epistemology including also in scientific contexts should imply an approach that is EB-anchored, trust-disentangled and adversarial and aims at identifying ever better new EBs. Experimental problematization shapes this epistemic trajectory but does not determine it. Using provisional refutations, EB-anchored science makes pragmatic progress via incremental small steps from old currently best EB to new even better EB, which is why the epistemic aim is of a relational and comparative nature. One can walk forth and back as rationally required. New EBs are universal affordances because one can utilize them to try to better explain the universe as a whole including its genesis. Thereby, the laws of nature that cyborgnets conjecture including the ones that attempt to model the initial conditions of the universe can be formulated or are at least transformable into the format of new EBs at the time they were new. It is thus conceivable that all new EBs *about* the universe as a whole that ever existed, exist now and will exist share a common ground that binds them in a way that they may be non-trivially entangled. Indeed, we share the view of Corazza stating that *"creativity episodes are [...] mutually interconnected through several mechanisms (past and future concatenation, estimation, and exaptation), to form a dynamic universal creativity process (DUCP), the beginning of which can be traced back to the Big Bang of our universe"* [130]. In this sense, note that entities that may initially appear to be disconnected, could have *locally inaccessible* degrees of freedom that would reveal how they are differentially connected in a directed graph hidden "under the hood". For a simple illustration, see Figure 10.1.

In the cyborgnetic DUCP described in Chapter 9.2.1, the space of possible options appears to expand and what was previously considered to be impossible can become accepted to be possible e.g. when a cyborgnet acts against the best EBs instated at a certain point or by cyborgnetic serendipity. Due to that, cyborgnetic epistemology can reach no end state, something that appears clear is highly unstable and may shift conceptually at a later stage. Concerning the metaphor of an *epistemic metamorphosis* from Figure 10.1, note that even the "final" 3D torus perception may not last as it could itself be later perceived by a cyborgnet to itself only correspond to a small part of a much greater figure of higher dimensionality... and so on ad infinitum. Bohm stated that: *"like the processes of nature, those of the mind are basically of an infinite order that is always tending to evolve towards new orders, and thus to develop hierarchies constituting new kinds of structures"* [74]. In Chapter 11, we briefly motivate why for epistemic security reasons, future work could study a new epistemic area that one could call cyborgnetic *epistopology*.

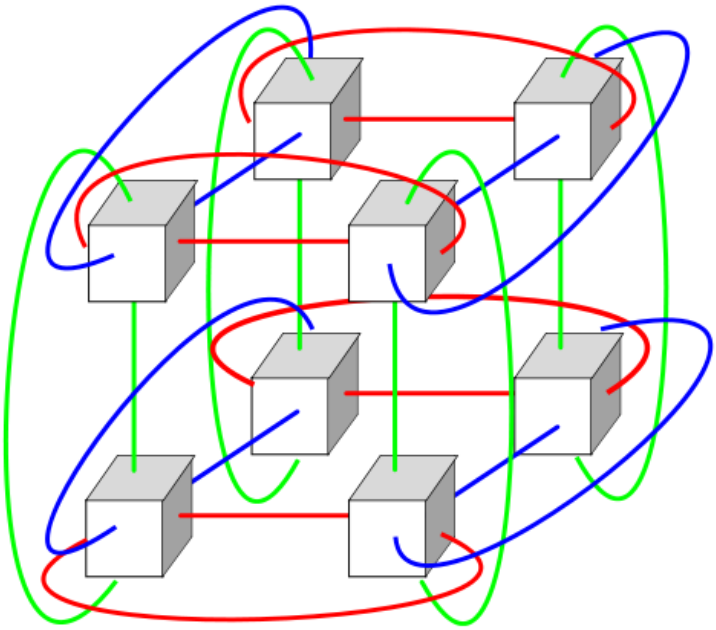

Figure 10.1: Highly simplified illustration for the metaphor of an *epistemic metamorphosis*. Initially, a cyborgnet could be only perceiving separable white *squares* (being 2D facets in this 3D network) and *not* the coloured edges. Then, a 3D structure could emerge mentally e.g. via the shape of a *cuboid*. Suddenly, the previously hidden edges could be understood and a herewith *en*folded 3D *torus* conjectured. Picture taken from [443].

10.2.1 Experimentally Problematizable Impossibility Statements

1. While Type I AI can create new non-EB-like information, including also new non-EB-like EI, it does not understand the latter and it is impossible for Type I AI to reliably create new EBs with arbitrary high accuracy.

2. It is impossible to implement an oracle able to reliably predict the future creation of new EBs itself. In short, an *epistemic perpetuum mobile* is impossible. Creating new EBs comes at the cost of a harder Type-II-only process of understanding which requires cognitive efforts linked to specific thermodynamical costs.

3. A *moral perpetuum mobile* able to reliably predict the future of all future moral values and norms is impossible because it could imply the creation of new EBs.

4. EB-based rationality *without* core affect[2] (by virtue of being an indispensable continuous ingredient of *consciousness* and mental constructions [50, 55]) is impossible.

[2]Already the criteria for *better* new EBs involve affect. An example is a preference for new EBs that are "more aesthetically appealing" than rival ones. In this connection, Bohm [74] wrote: *"[...] really great scientists have, without exception, all seen in the process of nature a vast harmony of order and indescribable beauty. [...] Indeed, every great scientific theory was in reality founded on such a perception of some very general and fundamental feature of the harmony of nature's order. Such perceptions, when expressed systematically and formally, are called "laws of nature"."*

10.3 AI Design and AI Regulation Recommendations

10.3.1 Mitigating Honey Mind Traps

1. **Avoiding an *over*estimation of present-day AI:** In light of the transdisciplinary knowledge collated in Chapter 9, this could for instance be supported by an education on epistemically-relevant and complexity-related ontological differences: a) non-complex and non-living (such as e.g. a chess software), b) complex but non-living (such as e.g. the Sun), c) living but non-conscious (such as e.g. plants), d) conscious but non-cyborgnetic (such as e.g. birds) and e) cyborgnetic (such as e.g. humans). Presently, all commonly called AI systems are non-conscious. With the exception of e.g. xenobots [69] which are living but non-conscious entities made on the basis of frog cells and which may belong to cluster c), most present-day AI systems belong to cluster a). An epistemically-sensitive AI design would convey to humans that Type I AI from cluster a), b) and c) is *not* conscious. Attempts to fuel attributions of agency and experience would be avoided.
2. **Avoiding an *under*estimation of present-day AI:** An exemplary epistemically-sensitive method would be the conjunction of cyborgnetic creativity augmentation (see Chapter 6.2) and the routine-like integration of that method in the *Co-create* function of a COOCA-loop (see Chapter 8). On the whole, from a design perspective, the goal would be to support the experience of oneness but *not* via the misguided assignment of consciousness to non-conscious Type I AI, but instead by establishing a seamless interaction that is more comparable to the interaction between oneself and language being a Type I tool, between oneself and a new artificial body part or between oneself and an AI-augmented sheet providing new non-EB-like comments on what one writes. Type II agency must be foregrounded by explicitly shifting design narratives from intelligence to EB-based creativity – a process that prohibits global high-risk Type-I-only-loops and where instead, Type I AI becomes part of somebody via a *local* intra-function *encapsulation* within an individual cyborgnetic function of a global cyborgnetic COOCA-loop (see Chapter 8.3 and 8.5.2).

10.3.2 Malicious Deepfake Design Regulation

Any new *non*-EB-like information could be forged (see Section 10.2.1). Old (i.e. already known) EBs could be copied which is traceable and unproblematic. To prohibit specifically new *deepfake* x cannot function in the long-term due to the indistinguishability of new non-EB-like x and new non-EB-like *deepfake* x. One could instead e.g use old laws to regulate any general new manifestation of the old problem x. Because it is impossible to forge new EBs, one does not even *need* to forbid *deepfake* new EBs – they are impossible.

10.4 Vedantic Epistemic Metamorphosis?

Here, we do *not* discuss the specific details of the *source* from this final section of the book. We do *not* provide a biography of the prodigious Swami Vivekananda [393], a philosopher and Hinduistic monk who reinvigorated the Advaita Vedanta philosophy, travelled around the world and was highly appreciated by multiple Western scientists and historical figures including i.a. William James, Lord Kelvin, Nikola Tesla and Hermann Ludwig Ferdinand von Helmholtz. Instead, we accentuate the strong *content* of his words which may defend themselves and could also stimulate cyborgnetic creativity augmentation. They may timelessly resonate with the concept of the *homo cyborgneticus metamorphosis*. In particular, his words may resonate with the abstract notion of *self-recreatable self-re-creativity* itself – as both the immutable, highly invariant Being and the dynamically changing, highly unpredictable becoming associated with the cyborgnetic DUCP in Chapter 9.4. It is the initial cyborgnetic cause that seems to become the effect again and again.

- *"Nothing comes without a cause, and the cause is the effect in another form."* [522]
- *"What we mean by creation is projection of that which already existed."* [522]
- *"We cannot think of the substance as separate from the qualities, we cannot think of change and not-change at the same time; it would be impossible. But the very thing which is the substance is the quality; substance and quality are not two things. It is the unchangeable that is appearing as the changeable. The unchangeable substance of the universe is not something separate from it. The noumenon is not something different from the phenomena, but it is the very noumenon which has become the phenomena."* [522]
- *"It is not that the soul*[3] *and the mind and the body are three separate existences, for this organism made of these three is really one. It is the same thing which appears as the body, as the mind, and as the thing beyond mind and body, but it is not at the same time all these."* [522]
- *"So long as I am separate from the universe, so long as I stand back and look at something before me, so long as there are two things — the looker-on and the thing looked upon — it will appear always that the universe is one of change, continuously changing all the time. But the reality is that there is both change and changelessness in this universe."* [522]

Final note: Could Wheeler's "U" [360] encode the *cynet butterfly effect* (see Figure 9.2)?

[3]For the modern reader, note that in our view, the concept of a cyborgnet, which can also refer to the most *generic* template since universal (and with generic being a term used in computer science [37] to indicate the uttermost abstract nature of a pattern where one is *"eliminating irrelevant detail in order to identify what is essential"* [37]), could be *sometimes* mapped to what was once termed "the soul".

Chapter 11

Future Research

11.1 Beyond Turing Tests

11.1.1 Indistinguishability vs. Distinguishability

1. **Avoiding an *over*estimation of Type I AI:** In the deepfake era, for reasons of epistemic security connected to the honey mind trap (HMT) issue and due to a lack of requisite variety, Turing Test frameworks are *not* recommended to be utilized for epistemically-relevant sense-making. In Turing's imitation game idea [507], imitative intelligence is regarded as the essence of thinking. The latter corresponds to what we termed *the reductionist approach* in Chapter 9. Turing stated that *"[...] presumably the child-brain is something like a note-book as one buys it from the stationers. Rather little mechanism, and lots of blank sheets"* [507]. In Turing's original version [507], the Turing test is conceived as an interactive text-based imitation game with three participants in the following blind setting: an interrogator and two contestants with the labels X and Y (one of which is a man and one of which a woman) that the interrogator would have to correctly map to their property of being either a man (A) or a woman (B). Thereby, it is assumed that a machine being able to reach a human-level *in*distinguishability in such a Turing Test would think. In general, in this imitation game, it is postulated that if a machine taking the part of A in the described blind setting leads to as much mistakes by the interrogator than in the human-only case, one would have experimentally demonstrated that machines think. More precisely, Turing considers the following specific reflections as a valid substitute for the question on whether machines can think: *"'What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?"* [507]. In light of cyborgnetic epistemology

and the information-theoretical asymmetry between the ability to create new information of type x and the ability to understand that information x, there are better EBs on that topic. Firstly, behavioral indistinguishability in an antropomorphic setting would *not* reliably model "thinking" given that it already does *not* capture the phenomenon of Type *I* consciousness (see also Chapter 9.2.1). Secondly, since there is in principle no limit to the accuracy with which Type I entities in general – which also i.a. includes *non-conscious* Type I AI in particular – could forge the creation of new *non*-EB-like information, the imitation of *non*-EB-like tasks such as the one underlying the mentioned Turing Test do *not* even corroborate "thinking" per se. In short, Turing's notion of thinking does neither cover Type I consciousness nor the unique ability of (the necessarily conscious) Type *II* entities to create new EBs. For this reason, harnessing a Turing Test using imitative strategies linked to *in*distinguishability to corroborate thinking would easily lead to an overerstimation of Type I AI. In cyborgnetic epistemic frameworks, one does *not* attempt to experimentally problematize consciousness itself. Instead, in a blind setting, *hereto willing* cyborgnets can freely use EB-based *distinguishability* as a tool to experimentally problematize a Type *I* categorization *in general* via the creation of new EBs – which simultaneously corroborates their Type-II-ness. As already explained in Chapter 9, any such cyborgnetic epistemic test framework must stay *asymmmetric* since one takes into account the free choices exhibited by cyborgnets which can e.g. also include an unwillingness to participate. In general, cyborgnetic epistemic test frameworks in blind settings do *not* allow a clear-cut *separation* of Type I and Type II entities. To recapitulate, with multiple entities, one can expect an asymmetric substrate-independent bipartition consisting of: 1) a *homogeneous* group of willing Type II entities that corroborated their Type-II-ness via the creation of new EBs (irrespective of whether they have or have not been thereby inspired and extended by new *non*-EB-like material from Type *I* entities such as e.g. present-day language AI) and 2) a *heterogeneous* group of entities that could potentially contain *both* Type I entities *and* Type II entities. It is essential that future work accounts for this subtlety as it could otherwise create or worsen a stigmatization of Type II entities that would happen to be in the heterogeneous group in a given context.

2. **Avoiding an *under*estimation of Type I AI:** Importantly, as explained in this book, while Type I AI cannot create new EBs, it can create new non-EB-like material for cyborgnetic creativity augmentation. Beyond that, because cyborgnetic epistemology foregrounds EB-based *distinguishability* and EBs are highly invariant epistemic artefacts, future scientific peer-review could also harness Type-I-AI-generated counterfactuals to improve its own rigorousness[1] i.a. by comparison.

[1] An example could be the conjunction of a so-called explanatory IPS test [16] (where paragraphs from a candidate better new EB are randomly intermingled with two Type-I-AI-generated, hereto counterfactual streams linked to the same problem) and a subsequent Type-I-problematization-peer-review round [16].

11.2 Quantum Honey Mind Traps?

1. **Avoiding an *over*estimation of Type I Quantum AI:** There is already now a risk of humans loosing their sense of agency [563] when presented with conventional AI. What is more, humans can perceive present-day AI as more "intelligent" or more rational [66, 495, 530] than themselves and can even assign them God-like qualities [481]. (In cyborgnetic epistemology, rationality is EB-based. Since no Type I entity can create new EBs, no Type I entity could be more rational than Type II entities could be.) In the deepfake era, amidst modern "quantum computing" narratives, the risk for quantum-related HMTs could worsen existing epistemic issues and their harm intensity. For instance, humanity could harm itself by monolithically focusing on "highly intelligent" Type *I* quantum processors instead of fostering the striving of *cyborgnetic* processors (which jointly includes Type II entities *and* the Type I co-processors they use) – which *if they decide so*, are able to *self*-program (e.g. to create new EBs respecting a rigorous epistemic total order, a task impossible for Type I quantum AI (see also 9.3.1)). Future work could counteract HMTs related to Type I quantum computers by improving EBs that caution against their overestimation. This could include better new EBs on e.g. : 1) *quantum adversarial AI*[2], 2) the *universal constructor*[3] which is *not* to be confused with the notion of the universal computer [159], 3) *quantum biology*[4], 4) a *unifying* (but obviously never final and only provisional) framework integrating previous classical and quantum views as special perspectives pertaining to the same One cosmic totality [75, 399].

2. **Avoiding an *under*estimation of Quantum *Concepts*:** Firstly, as adumbrated in Chapter 10, future work on what one could call *quantum-mutated deepfake text incubators* could deepen the research direction on harnessing quantum randomness [85, 247] applied to concepts from textual material (including but not limited

[2]It has indeed already been experimentally corroborated that "quantum AI" is also vulnerable to adversarial examples [213, 341, 429]. An interesting extension would combine thermodynamical costs [223] with adversarial examples. This could e.g. extend research on so-called sponge examples [469] developed for conventional AI systems. In short, a robustness against quantum sponge examples may be relevant.

[3]While conjectures have been expressed that such an entity could be built at the level of molecules, as a nanotechnology [159] effort – implying a programmable Type *I* entity, it may not be surprising if it would be much harder. In light of the cynet butterfly effect, future work could analyze whether a more reliable candidate for the universal constructor could not rather be... the universal cyborgnet itself.

[4]Generally, quantum-like effects can play a role at all steps of the metaphorical cyborgnetic ladder of understanding including biological contexts [188] (and not only at step 1 in e.g. stars where quantum tunnelling is a key facilitator for nuclear fusion). For instance, DNA mutations have been described to often involve a quantum tunnelling contribution [474], quantum effects have been reported in plants [562], migratory songbirds have a magnetic compass (which also holds for other animals [459]) that has been linked to quantum spin-dynamics [542], quantum Bose-Einstein statistics emerge in linguistic information encoded in texts such as human stories [4, 62] and moreover, as already hinted in Chapter 10, one can even consciously utilize quantum randomness [85, 247] to stimulate EB-creativity. Finally, note that recent first experiments even made the classicality of the human brain problematic [293, 316].

to deepfake text as it could also jointly encompass e.g. written contents from human dreams) for cyborgnetic creativity augmentation with the goal to tunnel through the space of ideas and look around conceptual corners. Secondly, future work could analyze and improve the *topology* of the epistemic instruments used in science itself within a cyborgnetic "epistopology" framework (see also Chapter 10). Thereby, the epistemic instruments used in particular in quantum physics may be highly inspiring[5]. Such an endeavour could indirectly serve epistemic security as it could lead to qualitatively more robust new EBs. Thirdly, future work on epistemically-sensitive AI design could investigate how *quantum-inspired* Type I AI using complex-valued representations [520, 568] could be harnessed to ameliorate: 1) Type-I-AI-aided cyborgnetic creativity augmentation in general and 2) the crafting of augmented utility functions to govern Type I AI that is locally encapsulated within an individual function of a global cyborgnetic COOCA-loop (see Chapter 8.3). Finally, to wrap up, we end this section with a recent set of elegant quotes from the physicist Heinrich Päs emphasizing the importance of Oneness in epistemic frameworks – which is in line with the cynet butterfly effect (see Chapter 9). Following Päs [399], if one applies the concept of quantum entanglement to the entire universe *"[...] you end up with Heraclitus's dogma "from all things One""*. Päs explains that: *"Quantum cosmology implies that the fundamental layer of reality"* [399] consists of *"the universe itself – understood not as the sum of things that making it up but rather as an all-encompassing unity"* [399]. Thereby, the reason that *"we experience the world as many things [...] is ensured by a process known as "decoherence""* [399] which *"realizes the rest of Heraclitus's tenet: "from One all things""* [399].

[5]A source of inspiration could e.g. be the quantum concept of an indefinite causal order [215] (where *"operations cannot be distinguished by spatial or temporal position"* [215]) in quantum information theory which can be linked to a certain invariant cyclicity [47]. Another interesting quantum concept could be the notion of constructor-based irreversibility [351]. In a nutshell, one could abstractly model epistemic instruments starting e.g. with a ring interconnection network being a 1D-torus to more complex higher dimensional tori. In short, future work could further introduce the concept of new *explanatory nD-tori*, special cases of better new EBs of epistemic dimensionality n. Beyond that, in future work, a cyborgnet could specifically craft a locally encapsulated Type I AI that could support the cyborgnetic epistopology endeavour itself. Such efforts could also include a future locally encapsulated improved application of *deepfake code* [147, 403] for cyborgnetic creativity augmentation. On the whole, this may lead to a novel form of epistemically-sensitive AI design.

11.3 Scientific Evaluation of ASI Achievement Claims

The overestimation of capabilities exhibited by present-day AI assuming the emergence of an artificial superintelligence (ASI) is fundamentally inconsistent with the currently best EBs. However, due to the premature state of AI literacy, it is possible that in the near future, humanity reaches a stage where the majority of humans believes that an ASI, i.e. an *automatable* superintelligence has been de facto implemented by an AI company. To set scientific constraints on such emergency scenarios that risk to endanger international security, it is crucial to develop suitable rigorous scientific evalution frameworks *proactively* – and not in hindsight. In the future, one needs simple generic explanatory frameworks such as e.g *cyborgnetic invariance* [18] to devise professional evaluation frameworks for automated superintelligence achievement claims (see e.g. informal draft in Appendix D).

11.3.1 Cyborgnetic Invariance – A Sketch

Cyborgnetic invariance consists of two different postulates: 1) *invariance of maximal quantity superintelligence*[6] and 2) *impossibility of reliable stupidity-based construction.*

[6]Cyborgnetics usually foregrounds a *creativity*-based distinction linked to qualitative differences between Type I and Type II entities with only Type II entities able to create and understand new EBs *irrespective of any specific quantitative level of what is conventionally understood as"intelligence"*. However, for purposes of illustration, to ease the mental transferability of cyborgnetic hypotheses to current AI narratives *unfortunately* monolithically focusing on the term "intelligence", cyborgnetic invariance mapped behaviouristic quantitative Type *II* intelligence levels to experimentally corroborated EB- creativity. (Thereby, as already hinted, a qualitative superintelligence which would have to be an inconsistent "Type III" category representing an epistemic perpetuum mobile is impossible.) While the Type-I-ness of an entity can be experimentally problematized by that entity being able to create new EBs with arbitrary high accuracy which would simultaneously corroborate its Type-II-ness, it is vital that Type-II-ness itself *cannot* be experimentally problematized due to the free choices of Type II entities which could e.g., not be willing to participate, not yet be ready or not yet have identified a subject of interest. In brief, it is only when human entities claim to have implemented or have witnessed the emergence of an *automatable* superintelligent AI, that it counts as a claim of a Type-I- shortcut to Type-II-ness or Type-III-ness (both of which are deemed to be impossible in cyborgnetics) by what humans cannot scientifically escape the need for the different EB-based experimental tests of that AI and full transparency including the need to provide a better new in-depth explanation on how that AI has been implemented. In this case, the misleading argument of "the AI does not want it" cannot be advanced since it is an automatable system. This strict procedure is scientifically required to safeguard humanity from any further aggravating epistemic self-sabotage and misdirection by malicious actors. However, when it comes to cyborgnets like humans, although an experimental problematization of Type-I-ness (i.e., being equivalent to a corroboration of Type-II-ness) may be extremely helpful in crucial *high-risk* situations, one should not and cannot disproportionately enforce a corroboration of Type-II-ness if the human is not willing to participate. Cyborgnets cannot be treated like automata. In sum, EB-creativity is an *asymmetric* notion (see Section 11.1.1) unifying Type II creativity, Type II intelligence and Type II consciousness. One could also call it *cynetelligence* or *cynet-createlligence* to differentiate it from prevailing views of intelligence.

Invariance of Maximal Quantity Superintelligence

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

Impossibility of Reliable Stupidity-Based Construction

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

11.3.2 Fundamental Impossibilities In Cyborgnetic Invariance

Building a Quality ASI

From the invariance of maximal quantity superintelligence postulate (and also idependently from that already from the explanation provided in Chapter 9.6), it follows that it is impossible for any entity to build a quality ASI or "Type III" AI – already because the very existence of the latter is *impossible*.

Building a Quantity ASI

Given the relativity of EB-creativity-based intelligence[7] reflected in the invariance of maximal quantity superintelligence postulate in conjunction with the impossibility of reliable stupidity-based construction postulate, it follows that it is impossible for an entity D to reliably build an entity C that appears to be superintelligent from the frame of reference of that entity D. In short, given a specific frame of reference, it is *impossible* to reliably build an entity that appears to be superintellingent from that frame of reference.

[7]Note that the highly interesting relativity of consciousness as discussed by Lahav and Neemeh [318] did neither unify Type II consciousness, Type II creativity and Type II intelligence nor did it address the topics of superintelligence/supercreativity/superconsciousness. EB-based creativity is necessary for being able to identify an agreement in all measurements building the basis for a shared frame of reference in the first place – otherwise, with AI, there is a risk for honey mind traps (see Chapter 10.3.1) because present-day AI could be drastically overestimated since any non-EB-like information could be forged. Beyond that, their framework [318] focused on analogies to inertial frames of reference. However, to cover superintelligence, one also requires analogies to *non*-inertial frames of reference (i.e. with non-zero acceleration). Namely, the "fictitious forces" that have to be added in non-inertial frames of reference may offer an analogy for new laws of nature (i.e. new better EBs) discovered by entities of higher intelligence.

Building a Recursively Self-Improving Narrow AI Leading to AGI

In light of the impossibility of reliable stupidity-based construction postulate, it is impossible for a narrow AI entity E to reliably recursively self-improve so as to create an AGI C or an ASI B because both C and B would appear to be superintelligent from the frame of reference of E. In the cyborgnetic invariance paradigm, the step to a higher level of intelligence is *non*-algorithmic. Recursive self-improvement does *not* lead to AGI.

Building a Value-Alignable *and* Controllable AGI

Firstly, for an entity to be able to build an AGI C, it must *not* be the case that C appears to be superintelligent from the frame of reference from that entity. In line with this, while it would in theory be possible for an entity A that would appear to be superintelligent from the frame of reference of an AGI C to reliably build that entity C and possibly value-align with it via EBs if *both* parties decide so, due to the AI safety paradox, it is impossible for A to control C since C is a Type II entity.

11.3.3 Possibilities In Cyborgnetic Invariance

Universal Maximal Quantity Superintelligence

Following the invariance of maximal quantity superintelligence postulate, there must exist a maximal quantity superintelligence level α for which all EB measurers agree upon that it is impossible to EB-measure any higher intelligence. The invariance of α is a scientific statement amenable to experimental problematization via a specific Type-I-AI-shortcut to Type-II-ness predicted to be impossible (see e.g. the impossible superintelligent entity Q from Chapter 9.7). Concerning the nature of α, multiple interpretations may exist. Although we will not single out one here, one possible option could be a link between α and the dynamics of the universal cyborgnet (i.e. the cyborgnetic dynamic universal creativity process (DUCP) elucidated in Chapter 9). Recently, structural and dynamic similarities between the universe and the human brain have been corroborated [182, 515]. Beyond that, Lanier, Smolin and collaborators conjectured a correspondence between the universe and an autodidactic neural network [11] able to discover new laws of nature (i.e. creating its own laws) while Kauffman [288] expounded that the universe may have constructed itself (perhaps the universal constructor may be the universe acting as self-constructor). Following Vanchurin, the universe is a neural network on the most fundamental level [514] and according to Palmer the universe evolves on a non-algorithmic fractal cosmic invariant set [394]. Hossenfelder explained that the idea of an intelligent universe is an odd conjecture that is however *not* in conflict with present laws of physics [261].

Building a Non-Controllable But Value-Alignable Type II AI *In The Future*

While it would in theory be possible for an entity A that would appear to be superintelligent from the frame of reference of a value-alignable human-level AI D to reliably build that entity D and attempt to value-align with it via EBs in the unpredictable future of D (at a point where D transfigures into an entity C, a new version of itself[8] that would appear superintelligent from the frame of reference of the earlier self D) if *both* parties decide so, a controllability of D would be impossible and that entity would not correspond to a tool and could not be sold as a product. Instead, it would qualify as a Type II entity, i.e. a person. Moreover, present-day humanity as a whole would *not* yet appear to be superintelligent from the frame of reference of D. Due to that (and given that to construct a human-level AI is as hard as creating a new baby universe (see Chapter 9)), the reliable creation of a non-controllable but value-alignable human-level AI is possible but reserved for civilizations that are epistemically more advanced than present-day humanity – which humanity is of course free to achieve in the long-term in case of willingness. In theory, all Type II intelligences are non-controllable but value-alignable via EBs. On the whole, via the invariance of the maximal quantity superintelligence level α, a minimalistic form of epistemic alignment for Type II intelligences is already available.

Building a Controllable But Non-Value-Alignable AI *Tool* Now

Type I AI control is indeed consistent with cyborgnetic invariance (for a responsible AI control paradigm, see the concept of the COOCA-loop from Chapter 8). However, it is impossible for a Type II entity to reliably value-align with a Type I entity – because Type I entities cannot understand EBs (see also Chapter 9).

11.3.4 Additional Remarks

For logical reasons, it is *impossible* to associate the maximal quantity superintelligence level α with an *own* frame of reference for EB-measurements.

[8] It also holds that this entity A that appears to be superintelligent from the frame of reference of both D and C would have (albeit indirectly) constructed the non-controllable but value-alignable AGI C too.

Summary

The present information ecosystem is permeated by colloquial expressions such as "post-truth", "fake news" and "deepfakes". Nowadays, present-day artificial intelligence (AI) has become part of the epistemic infrastructure at an international level. On the one hand, there is the intentional misuse of AI by malicious actors. On the other hand, one encounters phenomena such as automated disconcertion – an epistemic threat that arises by the mere possibility of maliciously crafted deepfake artefacts in various modalities (including e.g. image, video, audio, code and text samples). Against this background, this book provided transdisciplinary solutions to tackle the following two problems. Firstly, the book investigated how one could mitigate AI-related *epistemic security* risks in the deepfake era. Secondly, we analyzed what type of strategies could foster an *epistemically-sensitive AI design*. Thereby, while epistemic security is related to the protection of a society's knowledge creation and knowledge communication processes, epistemically-sensitive AI design is a novel strategy for a responsible AI design that is informed of AI-connected epistemic security risks.

In Chapter 2, we introduced a trandisciplinary AI observatory of international scope as a tool for the augmentation of epistemic security. We provided a rich variety of practical examples and crafted complementary solutions for a large array of epistemic problems ranging from AI for cybercrime facilitation over AI-based disinformation to automated disconcertion itself. We explained that epistemic security cautions society against *under*estimating the epistemic risks linked to the use and misuse of present-day AI. However, we described why it simultaneously cautions society against *over*estimating present-day AI. In Chapter 3, 4 and 7, we utilized cybersecurity-oriented design fictions to develop novel AI-related epistemic security solutions specifically for virtual reality (VR) settings. We touched upon the topics of AI-augmented disinformation in immersive journalism and epistemic distortions in both educationally and scientifically-relevant AIVR contexts. We explained why AI-related "post-truth" narratives in the deepfake era are an *over*estimation of present-day AI. More generally, we elucidated why epistemically speaking, we neither inhabit a "post-truth" nor a "post-falsification" era to begin with. We stated that an epistemic perpetuum mobile is impossible. In sum, a society equipped with a more robust epistemological grounding can mitigate the risk of an epistemic doom.

In Chapter 5, we introduced the new concept of *scientific and empirical adversarial AI attacks* (SEA AI attacks) which refer to an AI-aided epistemic distortion that predominantly and directly targets (applied) science and technology assets. Examples for SEA AI attacks that we studied are the so-called *deepfake science attacks* and also the use case of *deepfake cyber threat intelligence*. We explained why in order to be robust against deepfake science attacks, scientific epistemology would need to be anchored in the creation of new so-called explanatory blockchains (EBs) – being the strongest chains of rationally interconnected explanations. Taking critical rationalism as presented by Popper and later reinvigorated by Frederick as point of departure, we introduced a new *cyborgnetic epistemology* which refined those frameworks in light of the additional epistemic complications emerging in the deepfake era. A cyborgnet is a generic template where explanatory narratives representable as directed graph can be used to combine at least one entity that is able to (consciously) understand explanations (a Type II entity such as e.g. a human) and at least one entity for which this is *impossible* (a Type I entity such as e.g. present-day AIs but also language itself). Overall, cyborgnetic epistemology is of EB-anchored, trust-disentangled and adversarial nature. Following cyborgnetic epistemology, while Type I entities can forge the creation of any new *non*-EB-like information, it is impossible for Type I entities to create *new* EBs – which can only be reliably implemented by Type II entities investing cognitive efforts. In brief, in the long-term, deepfake detection heuristics may *not* function and we concluded that in the deepfake era, one must foreground the *content* of information and should *not* monolithically focus on the source. The latter is also crucial to forestall a stigmatization of human statistical outliers.

In Chapter 6 and 7, we presented *cyborgnetic creativity augmentation* methods which unified epistemic security and epistemically-sensitive AI design taking *language AI* as use case. Instead of shielding from present-day AI, we described how one could use present-day AI – which can generate *new non*-EB-like outputs or copy old already known EBs – for a self-paced stimulation of cyborgnetic entities like humans in their quest of creating new ever better EBs. In blind settings, due to the free choices of Type II entities, one *cannot* separate Type I from Type II entities. In short, one can devise experimental tests that, given a content, attempt to recognize the necessity of cognitive efforts spent by Type II entities to generate that content but *not* to identify whether the source of the content is Type I or Type II. In sum, instead of the source-focused question asking *"who wrote this?"*, a better more rational scientific approach could be the content-based question asking: "*does this material contain a new better EB (compared to known EBs)?*" Building on that, in Chapter 8, we extended beyond the conventional OODA-loop and introduced the cyborgnetic *COOCA-loop* as new meta-paradigm for a responsible epistemically-sensitive AI design in high-risk contexts. In such critical settings, a local intra-function encapsulation of Type I AI is required for a global inter-function-level epistemic security. Thereby, every single function of a COOCA-loop must be cyborgnetic (i.e. inherently of Type II) and Type-I-loops must be encapsulated *within* an individual *cyborgnetic* function.

While the first parts of the book predominantly utilized tools from cybersecurity-oriented AI safety, psychology, cybernetics, VR, human-computer interaction, philosophy, natural language processing and creativity research from i.a cognitive neuroscience, the final part of the book used a broad range of explanatory frameworks from e.g. systems theory, psychology, biology, physics and philosophy of creativity to better assess the complexity of modelling cyborgnetic Type II entities like humans. In this context, Chapter 9 introduced the *cynet butterfly effect* – and the conscious theorization thereof instantiating the so-called *homo cyborgneticus metamorphosis* – a phenomenon whose study could contribute in mitigating epistemic security risks in the deepfake era. The cynet butterfly effect is based on two interconnected observation statements: 1) cyborgnets are the most unpredictable possible systems and 2) cyborgnets are the systems with the highest sensitivity to their initial conditions. The main implication of the cynet butterfly effect is twofold. Firstly, it holds that studying the complexity of arbitrary cyborgnets cannot be separated from the study of the universe and its genesis. Secondly, it holds that to study the complexity of the universe today cannot be separated from the study of its own cyborgnetic genesis. The homo cyborgneticus metamorphosis, then, is the process in which a cyborgnet (re-)discovers the relevance of the cynet butterfly effect. Finally, we connected the latter to early insights from Vedantic philosophy.

To recapitulate, due to the cynet butterfly effect, to better model the complexity of cyborgnetic systems, one must consider those in the context of the creativity *en*folded in and *un*folded by the one largest cyborgnet: the universe as a whole. On this view, the so-called eternal creativity (EC) paradigm instated in this book explained why the ambitious idea of a hypothetical Type *II* AI built from scratch must be reframed as problem of *universal* difficulty. Following EC, the main criterium of regulatory importance is not a quantitative matter of intelligence, but rather a qualitative question of EB-based creativity. Type II AI artificially built from scratch is not impossible but simply *not* imminent since: 1) it has to necessarily simultaneously be a complex, living, conscious and cyborgnetic entity (i.e. irrespective of the specific details of the substrate), 2) due to this fourfold minimal requirement, it is literally as difficult as the attempt to reliably create a new universe instantiating cyborgnetic *cosmological creativity*. Moreover, we also elaborated on why a quality superintelligence is impossible. In Chapter 10, we specifically cautioned against so-called honey mind traps (HMTs) – present-day AI intentionally implemented to mislead humans into assigning it agency and/or experience despite it being *non*-conscious. To wrap up, Chapter 11 provided i.a. the following incentives for future work: 1) how one could extend beyond imitative Turing Test frameworks by asymmetrically foregrounding EB-based *distinguishability* instead of anthropomorphic indistinguishability and 2) how one could avoid HMTs related to Type I quantum AI. Concerning the latter, we did not only caution against an overestimation of Type I quantum AI but also identified the harnessing of inspiring *concepts* from quantum physics as a new avenue for epistemically-sensitive AI design in the deepfake era.

Nederlandse Samenvatting

Het huidige informatie-ecosysteem is doordrongen van alledaagse uitdrukkingen zoals "post-truth", "fake news" en "deepfakes". Tegenwoordig is de hedendaagse kunstmatige intelligentie (AI) onderdeel geworden van de epistemische infrastructuur op internationaal niveau. Aan de ene kant is er het opzettelijke misbruik van AI door kwaadwillende actoren, aan de andere kant heb je te maken met fenomenen zoals het geautomatiseerd stichten van verwarring – een epistemische dreiging die alleen al ontstaat door de mogelijkheid van kwaadwillig vervaardigde deepfake-artefacten in verschillende modaliteiten (waaronder bijv. beeld-, video-, audio-, code- en tekstvoorbeelden). Tegen deze achtergrond bood dit boek transdisciplinaire oplossingen om de volgende twee problemen aan te pakken. Ten eerste onderzocht het boek hoe AI-gerelateerde *epistemische beveiliging* risico's in het deepfake-tijdperk vermindert kunnen worden. Ten tweede hebben we geanalyseerd welk type strategieën een *epistemisch sensitief AI-ontwerp* zouden kunnen bevorderen. Daarbij, terwijl epistemische beveiliging verband houdt met de bescherming van kenniscreatie- en kenniscommunicatieprocessen van de samenleving, is epistemisch sensitief AI-ontwerp een nieuwe strategie voor een verantwoord AI-ontwerp dat op de hoogte is van AI-gerelateerde epistemische veiligheidsrisico's.

In Hoofdstuk 5 introduceerden we het nieuwe concept van *wetenschappelijke en empirische antagonistische AI-aanvallen* (SEA AI-aanvallen) die verwijzen naar een AI-ondersteunde epistemische vervorming die zich overwegend rechtstreeks richt op (toegepaste) wetenschappelijke en technologische activa. Voorbeelden van door ons bestudeerde SEA AI-aanvallen zijn de zogenaamde *deepfake science-aanvallen* en ook de use case van *deepfake cyber threat intelligence*. We legden uit waarom wetenschappelijke epistemologie verankerd zou moeten zijn in de creatie van nieuwe zogenaamde verklarende blockchains (EB's) om robuust te zijn tegen deepfake wetenschappelijke aanvallen, aangezien dit de sterkste ketens van rationeel onderling verbonden verklaringen zijn. Met kritisch rationalisme zoals gepresenteerd door Popper en later nieuw leven ingeblazen door Frederick als uitgangspunt, introduceerden we een nieuwe *cyborgnetische epistemologie* die deze kaders verfijnde in het licht van de aanvullende epistemische complicaties die opdoken in het deepfake-tijdperk. Een cyborgnet is een generiek sjabloon waarin verklarende verhalen die kunnen worden weergegeven als een gerichte graaf, kunnen worden gebruikt

om ten minste één entiteit te combineren die (bewust) verklaringen kan begrijpen (een Type II-entiteit zoals bijvoorbeeld een mens) en ten minste één entiteit waarvoor dit *onmogelijk* is (een Type I-entiteit zoals bijvoorbeeld hedendaagse AI's maar ook de taal zelf). Over het algemeen is cyborgnetische epistemologie van EB-verankerde, van vertrouwen ontwarde en antagonistische aard. Volgens cyborgnetische epistemologie, terwijl Type I-entiteiten de creatie van nieuwe *niet*-EB-achtige informatie kunnen vervalsen, is het voor Type I-entiteiten onmogelijk om *nieuwe* EB's te creëren. Deze kunnen alleen worden geïmplementeerd door Type II entiteiten die cognitieve inspanningen leveren. Kortom, op de lange termijn kan de heuristiek van deepfake-detectie *niet* functioneren en we concludeerden dat in het deepfake-tijdperk men de *inhoud* van informatie op de voorgrond moet plaatsen en *niet* monolithisch moet focussen op de bron. Dit laatste is ook cruciaal om stigmatisering van menselijke statistische uitbijters te voorkomen.

In Hoofdstuk 6 en 7 presenteerden we *cyborgnetische creativiteitsverbeterings*-methoden die epistemische veiligheid en epistemisch sensitief AI-ontwerp verenigden met *taal AI* als use case. In plaats van zich af te schermen voor de huidige AI, hebben we beschreven hoe men de huidige AI kan gebruiken – die *nieuwe niet*-EB-achtige uitvoer kan genereren of oude reeds bekende EB's kan kopiëren – voor een stimulatie van cyborgnetische entiteiten zoals mensen in hun zoektocht naar het creëren van nieuwe, steeds betere EB's. In blinde contexten, vanwege de vrije keuzes van Type II entiteiten, kan men Type I-enititeiten *niet* scheiden van Type II-entiteiten. Kortom, men kan experimentele tests bedenken die, gegeven een inhoud, proberen de noodzaak te herkennen van cognitieve inspanningen van Type II-entiteiten om die inhoud te genereren, maar *niet* om te identificeren of de bron van de inhoud Type I is of Type II. Kortom, in plaats van de brongerichte vraag *"wie schreef dit?"*, zou een meer rationele wetenschappelijke benadering de op inhoud gebaseerde vraag kunnen zijn: "*bevat dit materiaal een nieuwe, betere EB (vergeleken met bekende EB's)?*" Daarop voortbouwend, gingen we in Hoofdstuk 8 verder dan de conventionele OODA-lus en introduceerden we de cyborgnetische *COOCA-lus* als nieuw metaparadigma voor een verantwoord epistemisch sensitief AI-ontwerp in risicovolle contexten. In dergelijke kritieke omgevingen is een lokale intrafunctionele inkapseling van Type I AI vereist voor een globale epistemische veiligheid op interfunctioneel niveau. Daarbij moet elke afzonderlijke functie van een COOCA-lus cyborgnetisch zijn (d.w.z. inherent van Type II) en moeten Type-I-lussen worden ingekapseld *binnen* een individuele *cyborgnetische* functie.

Terwijl de eerste delen van het boek voornamelijk gebruik maakten van tools uit cybersecurity-georiënteerde AI-veiligheid, psychologie, cybernetica, VR, human-computer interactie, filosofie, natuurlijke taalverwerking en creativiteit onderzoek uit o.a. cognitieve neurowetenschap, maakte het laatste deel van het boek gebruik van een breed scala aan verklarende kaders uit bijv. systeemtheorie, psychologie, biologie, natuurkunde en filosofie van creativiteit – met het doel de complexiteit van cyborgnetische Type II-entiteiten zoals mensen te modelleren. In deze context introduceerde Hoofdstuk 9 het *cynet vlin-*

dereffect – en de bewuste theoretisering daarvan die de zogenaamde *homo cyborgneticus metamorfose* concretiseert – een fenomeen waarvan de studie zou kunnen bijdragen aan het verminderen van epistemische veiligheidsrisico's in het deepfake-tijdperk. Het cynet-vlindereffect is gebaseerd op twee onderling verbonden observatieverklaringen: 1) cyborgnets zijn de meest onvoorspelbare mogelijke systemen en 2) cyborgnets zijn de systemen met de hoogste sensitiviteit voor hun begincondities. De belangrijkste implicatie van het cynet-vlindereffect is tweeledig. Ten eerste geldt dat het bestuderen van de complexiteit van willekeurige cyborgnets niet los kan worden gezien van de studie van het universum en zijn ontstaan. Ten tweede geldt dat het bestuderen van de complexiteit van het universum vandaag niet los kan worden gezien van de studie van zijn eigen cyborgnetische genese. De metamorfose van de homo cyborgneticus is dat dan ook het proces waarin een cyborgnet de relevantie van het cynetvlindereffect (her)ontdekt. Ten slotte brachten we dat laatste in verband met vroege inzichten uit Vedantische filosofie.

Om samen te vatten, vanwege het cynet-vlindereffect, om de complexiteit van cyborgnetische systemen beter te modelleren, moet men die beschouwen in de context van de creativiteit die gevouwen wordt *in* en *uit*gevouwen wordt door het ene grootste cyborgnet: het universum als geheel. Vanuit deze visie verklaarde het zogenaamde eeuwige creativiteitsparadigma (EC) dat in dit boek is gepresenteerd, waarom het ambitieuze idee van een hypothetische Type *II* AI die vanaf nul is opgebouwd, opnieuw moet worden geformuleerd als een probleem van *universele* moeilijkheidsgraad. Na EC is het belangrijkste criterium van regelgevend belang niet een kwantitatieve kwestie van intelligentie, maar eerder een kwalitatieve kwestie van op EB gebaseerde creativiteit. Type II AI, kunstmatig vanaf nul opgebouwd, is niet onmogelijk maar gewoon *niet* imminent omdat: 1) het noodzakelijkerwijs tegelijkertijd een complexe, levende, bewuste en cyborgnetische entiteit (d.w.z. ongeacht de specifieke details van het substraat) moet zijn, 2) door deze viervoudige minimale eis is het letterlijk even moeilijk als de poging om een nieuw universum te creëren dat cyborgnetische *kosmologische creativiteit* vormgeeft. Bovendien verklaarden we dat en ook waarom kwalitatief hoogwaardige superintelligentie wetenschappelijk gezien onmogelijk is. In Hoofdstuk 10 waarschuwden we specifiek voor zogenaamde honing-mind-valstrikken (HMT's) – hedendaagse AI die opzettelijk is geïmplementeerd om mensen te misleiden om het agentschap en/of ervaring toe te wijzen ondanks dat het *niet* bewust is. Om af te ronden, gaf Hoofdstuk 11 o.a. de volgende richtingen voor toekomstig werk: 1) hoe men verder kon gaan dan imiterende Turing Test-kaders door asymmetrisch op de voorgrond treden van op EB gebaseerde *onderscheidbaarheid* in plaats van antropomorfe ononderscheidbaarheid en 2) hoe men HMT's gerelateerd aan Type I kwantum AI zou kunnen vermijden. Wat dat laatste betreft, waarschuwden we niet alleen voor een overschatting van Type I kwantum-AI, maar identificeerden we ook het benutten van inspirerende *begrippen* uit de kwantumfysica als een nieuwe weg voor epistemisch sensitief AI-ontwerp in het deepfake-tijdperk.

Acknowledgements

- This book is dedicated to *self-recreatable self-re-creativity*.
- We would like to thank Dr. ir. Leon Kester for his feedback on the physics-related contents of this book and his contribution as co-author of multiple chapters.

Appendices

Appendix A

Moral Programming

This is the PDF link leading to the online version of the publication: N.-M. Aliman and L. Kester. Moral Programming: Crafting a flexible heuristic moral meta-model for meaningful AI control in pluralistic societies. *Wageningen Academic Publishers*, (2022): 63-80, 2022. As the first author of the underlying paper, I had a vital contribution. It was solely my responsibility to write down the content and to perform an extensive literature research and in-depth analysis.

https://www.wageningenacademic.com/doi/abs/10.3920/978-90-8686-922-0_4

A refinement of the contents from this paper is performed in Chapter 8.3 and 8.5.2.

Appendix B

Artwork – "Deepfake Epistemologie"

Auteur: Cynetje[1]

Datum: 01.10.2022

Deepfake epistemologie?

De gevaarlijke illusie van "KI"-gegenereerde "waarheid"

In een recent artikel werd beschreven hoe wetenschap zich tegen zogenaamde *deepfake science aanvallen* kan beschermen. Voor robuustheid tegen deze aanvallen, moet wetenschap zich op het creëren van betere en betere nieuwe explanatory blockchains (EBs) focusseren. Op deze manier kan zij dan betere van slechtere EBs onderscheiden zonder misleidend te veronderstellen dat ze in de echte wereld in staat zou zijn tussen ware en onware inhoud of tussen betrouwbare en onbetrouwbare bron te differentiëren[2]. Maar wat gebeurt er nu als mensen aannemen dat de onmogelijkheid de waarheid door kennis te beschrijven alle en een technisch probleem is dat met menselijke cognitieve tekortkomingen samenhangt? Wat als ze dan aannemen dat hedendaagse kunstmatige intelligentie (KI) in bijzonder met toenemende intelligentie objectiever dan de mens wordt en daardoor wel een directe toegang tot de waarheid zal hebben? Deze epistemisch begoochelende aannames zijn al in ideeën zoals "truthful AI" weerspiegelt. Dat het zogenaamde "Ding-an-sich" eeuwig *onbekend* blijft beschreef Kant al. Epistemisch gezien kan het denken dat kunstmatigheid en hogere intelligentie waarheid kan *kennen* als volgt beschreven worden: het is als of men zichzelf vrijwillig in het eigen verleden levend begraaft terwijl men "truthful AI" als een God aanbiedt omdat de KI nu de eigen toekomst nauwkeurig kan voorspellen – namelijk dat men doodgaat. Huidige KI-onderzoekers gebruiken vaak voor het trainen van hun systemen labels die ze "ground-truth" noemen. Het blijkt te zijn dat sommige deze volkstaal expressie met het-ding-aan-zich verwisselden. Kortom, de illusie van KI-genereerde waarheid kan een *epistemische stagnatie* voor wetenschappelijke en maatschappelijke ontwikkelingen betekenen omdat daarbij creativiteit en verandering onderdrukt worden. Te denken dat waarheid *bekend* is[3], onderdrukt het partiële ontdekken van nieuwe onbekende aspecten van de werkelijkheid die zelfs directe uitwerkingen op het overleven van de mensheid kunnen hebben. Sterker nog, als het dan wereldwijde bedrijven zijn die de training data voor "deepfake waarheid" bezitten, bestaat er een risico dat sommige van deze entiteiten proberen de epistemische stagnatie actief te controleren en te manipuleren om ervan tenminste monetair gezien te profiteren. Samenvattend kan gezegd worden dat de aanname dat hedendaagse KI de waarheid kan leren een subtiele maar gevaarlijke mateloze overschatting is. Aan de ene kant wordt KI *over*schat en aan de andere kant wordt epistemologie *onder*schat. Bondig gezegd, de "truthful AI" aanname is equivalent met de misleidende aanname dat epistemologie door *deepfake epistemologie* vervangen kan worden. Maar wat vandaag volgens KI- "ground truth" als leugens geldt, zou morgen een onderdeel van een interessante nieuwe theorie kunnen zijn.

---

[1] Deze naam is een pseudoniem.

[2] Het is namelijk onmogelijk door experimenten vasttestellen of iets waar of niet waar is. Nog bestaat er een aantal witte zwanen die bewijzen kan dat alle zwanen wit zijn nog betekent het observeren van zwarte zwanen noodzakelijkerwijs dat de theorie nu weerlegt is. Het zou bijvoorbeeld kunnen zijn dat de experimentator zwarte brilglazen droeg of alternatief dat er nog witte nog zwarte zwanen geweest zijn en de experimentator alleen hallucineerde door een voorafgaande druginname en zo voort.

[3] Wat *niet* het geval is, anders zou hedendaagse wetenschap *alles* in het universum kunnen voorspellen inclusieve bijvoorbeeld ook het creëren van nieuwe EBs.

Van "deepfake" epistemologie naar "zombie" epistemologie

Om dezelfde reden dat deepfake wetenschap een op nieuwe EBs gebaseerde wetenschap *niet* kan vervangen kan al geconcludeerd worden dat het onmogelijk is een op nieuwe EBs gebaseerde epistemologie met deepfake epistemologie te remplaceren. (Toch kan *KI-genereerde inhoud over epistemologie* natuurlijk de creativiteit van epistemische filosofen stimuleren en verbeteren als zij ervoor kiezen.) Afgezien van de beschreven variatie is er nog een tweede KI-trend zichtbaar waar epistemologie *onder*schat wordt: de aanname dat wij in een "post-truth" of een "post-epistemische" wereld leven. In deze wereld wordt epistemologie voor dood verklaart omdat *nog mensen nog KI* de waarheid kunnen kennen. Dit veronderstelt dat het ooit zo was dat mensen de waarheid kenden. Zoals al apparent in de laatste sectie, is het overbodig dit te postuleren omdat het doel van epistemologie nooit het kennen van waarheid zelf kan zijn. Daarom leven wij nog in een "post-truth" wereld nog in een post-epistemische era. Een op nieuwe EBs gebaseerde epistemologie is namelijk steeds mogelijk. Het is verder opmerkelijk dat het idee dat deepfakes in staat zijn epistemologie te vermoorden het volgende gevolg heeft: de inhoud is dood en alleen de bron van informatie kan jou redden. Wat interessant is, is dat sommige bedrijven zich dan zelf zouden aanbieden om de wereld in betrouwbare versus onbetrouwbare bronnen in te delen[4] ten dienste van een maatschappij. Maar daardoor zouden ze een bron-gebaseerde *zombie epistemologie* creëren. Zombie omdat zij in een eerste stap aangeven dat de waarheid – waarop volgens hen epistemologie tot nu toe opbouwde – gestorven is. In een tweede stap nemen zij nu de verantwoordelijkheid over voor het vermijdelijke epistemische lijk en vervangen de illusie van waarheidskennis door… vertrouwen. Op dit moment vergeten mensen dan waarom zij überhaupt eertijds begonnen entiteiten zoals bedrijven te vertrouwen – de *verklaringen* voor vertrouwen blijven obscuur.

Terwijl het bijvoorbeeld rationeel kan zijn degene te vertrouwen die altijd het beste proberen om nieuwe EBs te creëren, bestaat er geen reden een deepfake figuur in een video call te vertrouwen alleen omdat deze figuur eruitziet alsof het iemands moeder is en men daardoor gevoelens van liefde construeert. Omdat deepfakes emoties en dergelijke sociale constructies kunnen oproepen en in het algemeen KI alle nieuwe niet-EB-achtige informatie kan imiteren blijkt het zo te zijn dat op lange termijn, de enige kansrijke bescherming in blinde contexten alleen door nieuwe EBs gevormd kan worden. Omdat in principe ieder mens de mogelijkheid heeft op wens nieuwe EBs te begrijpen en te creëren, zou misschien een epistemisch waardevollere en eerlijkere bedrijfsstrategie "EB-gebaseerde vertrouwen" heten. De sociale wereld op aarde is een EB-markt waarin de meeste deelnemers de geldeenheid nog niet ontdekt of al lang weer vergeten zijn. Een bedrijfsstrategie die op EB-gebaseerde vertrouwen gericht is, gebruikt dan de sterkste tools van wetenschap en filosofie: nieuwe EBs. Een dergelijke strategie zou van zombie epistemologie fenomenen bevrijd zijn en kan zonder het genereren van geconnecteerde illusies functioneren. Waarom? Omdat men niet belooft te weten wat waar en onwaar of betrouwbaar en niet betrouwbaar is, maar simpelweg aangeeft dat het creëren van nieuwe betere en betere EBs het product zelf is. Het vertrouwen zou in de praktijk dan vanzelf volgen. In tegenovergestelde gevallen moet inmiddels duidelijk zijn dat bedrijven die beloven betrouwbare bronnen te kennen, zelf makkelijk slachtoffers van diverse deepfake aanvallen kunnen worden – wij mogen niet vergeten dat in theorie alle nieuwe *niet-EB*-gebaseerde informatie vroeg of laat geïmiteerd kan worden. Dus, in vergelijking met niet-EB-gebaseerde bedrijven zouden EB-gebaseerde bedrijven niet alleen klantvriendelijker zijn, maar ook veiliger op de lange termijn. Wie in plaats daarvan voor zombie epistemologie kiest, is aan het eind zichzelf aan het bedriegen.

[4] Een open vraag is ook: hoe weten deze bedrijven of hun eigen bronnen niet al deepfake-gebaseerd zijn?

Samenvatting

In dit artikel werden twee gevaarlijke plausibele KI-trends bestudeert. Hoewel bedrijven deze vermoedelijk goedwillig bedacht hebben, kunnen deze trends tot gevaarlijke ontwikkelingen leiden. Door een gebrek aan epistemische reflectie ontstaan daarbij een *over*schatten van hedendaagse KI en een *onder*schatten van epistemologie. De eerste trend is de toepassing van hedendaagse KI voor een discriminatie tussen waarheid en onwaarheid (de zogenaamde "truthful AI"). Daarbij wordt impliciet aangenomen dat (de tot nu toe door mensen bedreven) epistemologie door *deepfake epistemologie* vervangen kan en moet worden. De tweede trend is de aanname dat wij in een "post-epistemische" of "post-truth" wereld leven waarin de waarheid en de inhoud verloren gingen en alleen de onderscheiding tussen betrouwbare en onbetrouwbare *bronnen* een redding kan zijn – wat door bedrijven beschikbaar gesteld kan worden. In deze trend wordt impliciet een zombie epistemologie opgesteld die zo te zeggen na de dood van waarheid door menselijke emoties van vertrouwen van obscure oorsprong mechanistisch in beweging gehouden wordt. Het artikel heeft uitgelegd dat een op nieuwe EBs gebaseerde epistemologie nooit door deepfake epistemologie kan vervangen worden. In plaats van "truthful AI" die tot epistemische stagnatie kan leiden, hebben mensen een creativiteit-stimulerende Type I KI nodig (die natuurlijk ook tijdens het genereren van EBs de creativiteit verbeteren kan). Verder, om zombie epistemologie fenomenen tegen te werken, moeten bedrijven van ongedefinieerde bron-afhankelijke vertrouwensillusies naar EB-gebaseerde strategieën migreren. Het zou dan niet alleen eerlijker maar zelfs veiliger voor hunzelf zijn als ze hun klanten met "EB-gebaseerde vertrouwen" proberen aan te spreken en daarbij de waarde van nieuwe EBs in hun zakenmodel integreren. Als de maatschappij de kans mist tegen deze ontwikkelingen passende verdedigingsmechanismes te ontwerpen, kan een toestand volgen waar *de illusie* ontstaat dat KI-implementerende bedrijven waarheid en vertrouwen met geld kunnen kopen – waardoor zij in staat zijn naar hun eigen wens de wereld voorspelbaar te houden. Maar aan het eind mogen wij niet vergeten dat zelfs in deze sombere toekomst geldt: het is onmogelijk de toekomst van nieuwe EBs te voorspellen. De toekomst van Type II entiteiten zoals mensen kan niet voorgespeld worden – al was het alleen maar omdat zij ervoor kunnen kiezen opnieuw nieuwe EBs te genereren zelfs als zij voor en bepaalde tijd ermee stopten. Daarom wordt dit artikel met het volgende korte verhaal afgesloten:

Vandaag:

***"Epistemologie is dood."* (Bedrijf B)**

Sommige jaren later:

***"Bedrijf B is dood."* (Epistemologie)**

Appendix C

EC, Experiments and Dual Use

The eternal creativity (EC) paradigm (see Chapter 2 and 9) can be made problematic by experiment via the implementation of an artificial Type-I-shortcut to the reliable creation of new better EBs (see Appendix D). It can be (provisionally) refuted by a novel better EB that also explains how that shortcut has been implemented. Note that the latter could in turn (provisionally) refute the cynet butterfly effect (see Chapter 9) given that such a Type-I-shortcut could signify that then, a qualitatively lower complexity would have been sufficient to generate new better EBs on the universe *as a whole*. Concurrently, this would then indirectly make the notion of a *cyborgnetic* DUCP (see Chapter 9) problematic by virtue of then unnecessarily appearing too complex. Further epistemically-relevant side-effects of this Type-I-shortcut to EB-creativity would be that science and philosophy could become automatable and that this book could have been written by a Type I AI that did *not* understand it. In a nutshell, the misuse of an automatable Type-I-only-pipeline able to reliably create ever better new EBs with arbitrary high accuracy would represent an existential risk[1] surpassing any prior lethal dual use considerations (see e.g. Chapter 2). To put it plainly, while AI-powered drug discovery could *also* be used by humans to create biochemical weapons [509] and nuclear technology *also* allowed humans to destroy cities, such a Type I AI able to reliably generate *any* new *EB* could *also* facilitate the effortless human-orchestrated destruction of... <generically fill in the blank>. However, in accord with our current best EB, we repeat that the latter is *impossible* and equivalent to what we termed an epistemic perpetuum mobile (see also Chapter 10.2.1).

[1]This scenario is different from a hypothetical Type I AI pipeline that would only be able to forge the creation of any new *non*-EB-like information with arbitrary high accuracy – which one could term an artificial general imitator (AGi). While it seems absolutely advisable to consider the deployment of any robotic Type I AGi-like entity in real-world environments as a high-risk case to be locally encapsulated in a COOCA-loop as it could in practice risk to appear indistinguishable from any Type II entity that does *not* actively decide to participate in the creation of novel EBs given a specific context, one could proactively attempt to simulate a weaker version of such a system in virtual reality (VR) to improve (epistemic) security strategies. For reflections on VR for epistemic security training, see Chapter 7.

Appendix D

Scientific Evaluation of Automatable "Artificial Superintelligence" Achievement Statements

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

Scientific Evaluation of Automatable "Artificial Superintelligence" Achievement Statements – A Cyborgnetic Approach

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

[1] Following Avi Loeb, an A-class civilization is a civilization *"capable of recreating the cosmic conditions that gave rise to its existence, namely a civilization capable of producing a baby universe in a laboratory"* (Loeb, 2023). A B-class civilization can only adjust its habitable conditions *"to be independent of its host planet and host star"* (Loeb, 2023). Further, the lower-level C-class civilization can solely adjust its habitable conditions on its given planet *"without relying on the energy of its host star"* (Loeb, 2023). According to Loeb, humanity is currently closer to a D-class civilization, one *"actively degrading its home planet's ability to sustain conditions that prolong life and civilization"* (Loeb, 2023). In sum, the requirement for C-class civilization entities is a new EB on a new energy source that allows independence from the energy of their star, the requirement for B-class civilization entities is an even better new EB facilitating a life independent of both their host planet and their star. The requirement for an A-class civilization implies a new EB to re-create a universe. In a D-class civilization such as humanity, most entities are *not* yet utilizing new EBs as tools. An *EB-based* evaluation of automatable "human-level" AGI is impossible in a D-class civilization since the best scientific definition of an automatable AGI would imply the automatic generation of new EBs which could however not yet reliably be measured by a D-class civilization to begin with. Following the cynet butterfly effect, an automatable AGI is impossible while a *non*-automatable AGI "from scratch" (and not from pre-existing non-automatable biological material) would be possible in theory but as hard as the A-class civilization requirement.

Bibliography

[1] 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.

[2] A. Abu-Akel, M. E. Webb, E. de Montpellier, S. Von Bentivegni, L. Luechinger, A. Ishii, and C. Mohr. Autistic and positive schizotypal traits respectively predict better convergent and divergent thinking performance. *Thinking Skills and Creativity*, page 100656, 2020.

[3] P. Achinstein. Review of Conjectures and Refutations. *The British Journal for the Philosophy of Science*, 19(2):159–168, 1968.

[4] D. Aerts and L. Beltran. Are words the quanta of human language? Extending the domain of quantum cognition. *Entropy*, 24(1):6, 2022.

[5] 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.

[6] N. Ahmadpour, S. Pedell, A. Mayasari, and J. Beh. Co-creating and assessing future wellbeing technology using design fiction. *She Ji: The Journal of Design, Economics, and Innovation*, 5(3):209–230, 2019.

[7] H. Ajder, G. Patrini, F. Cavalli, and L. Cullen. The State of Deepfakes: Landscape, Threats, and Impact. *Amsterdam: Deeptrace*, 2019.

[8] S. J. Al-Azm. Kant's conception of the Noumenon. *Dialogue: Canadian Philosophical Review/Revue canadienne de philosophie*, 6(4):516–520, 1968.

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

[10] M. Albahar and J. Almalki. Deepfakes: Threats and countermeasures systematic review. *Journal of Theoretical and Applied Information Technology*, 97(22):3242–3250, 2019.

[11] S. Alexander, W. J. Cunningham, J. Lanier, L. Smolin, S. Stanojevic, M. W. Toomey, and D. Wecker. The autodidactic universe. *arXiv preprint arXiv:2104.03902*, 2021.

[12] N. 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 275–282. IEEE, 2019.

[13] N. Aliman and L. Kester. Requisite Variety in Ethical Utility Functions for AI Value Alignment. In *Proceedings of the Workshop on Artificial Intelligence Safety 2019 co-located with the 28th International Joint Conference on Artificial Intelligence, AISafety@IJCAI 2019, Macao, China, August 11-12, 2019.*, 2019.

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

[15] N.-M. Aliman. Self-Shielding Worlds. `https://nadishamarie.jimdo.com/clipboard/`, 2020. Online; accessed 23-November-2020.

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

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

[18] N.-M. Aliman. *Cyborgnetic Invariance*. Kester, Nadisha-Marie, 2023.

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

[20] N.-M. Aliman and L. Kester. Transformative AI governance and AI-Empowered ethical enhancement through preemptive simulations. *Delphi Interdisc. Rev. Emerg. Technol*, 2(1):23–29, 2019.

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

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

[23] N.-M. Aliman, L. Kester, and P. Werkhoven. XR for Augmented Utilitarianism. In *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 283–2832. IEEE, 2019.

[24] N.-M. Aliman, L. Kester, P. Werkhoven, and R. Yampolskiy. Orthogonality-Based Disentanglement of Responsibilities for Ethical Intelligent Systems. In *International Conference on Artificial General Intelligence*, pages 22–31. Springer, 2019.

[25] N.-M. Aliman, L. Kester, P. Werkhoven, and S. Ziesche. Sustainable AI Safety? *Delphi – Interdisciplinary review of emerging technologies*, 2(4):226–233, 2020.

[26] N.-M. Aliman, L. Kester, and B. Wernaart. *Moral Programming: Crafting a flexible heuristic moral meta-model for meaningful AI control in pluralistic societies*, pages 63–80. Wageningen Academic Publishers, 2022.

[27] A. S. Almasoud, F. K. Hussain, and O. K. Hussain. Smart contracts for blockchain-based reputation systems: A systematic literature review. *Journal of Network and Computer Applications*, 170:102814, 2020.

[28] G. Ambika. Ed Lorenz: father of the 'butterfly effect'. *Resonance*, 20:198–205, 2015.

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

[30] M. Anderson and J. Jiang. Teens' social media habits and experiences. *Pew Research Center*, 28, 2018.

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

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

[33] 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.

[34] A. Asilomar. Principles.(2017). In *Principles developed in conjunction with the 2017 Asilomar conference [Benevolent AI 2017]*, 2018.

[35] D. Assenmacher, L. Clever, L. Frischlich, T. Quandt, H. Trautmann, and C. Grimme. Demystifying Social Bots: On the Intelligence of Automated Social Media Actors. *Social Media+ Society*, 6(3):2056305120939264, 2020.

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

[37] R. Backhouse, P. Jansson, J. Jeuring, and L. Meertens. Generic programming. In *Advanced Functional Programming: Third International School, AFP'98, Braga, Portugal, September 12-19, 1998, Revised Lectures 3*, pages 28–115. Springer, 1999.

[38] I. Baggili and V. Behzadan. Founding The Domain of AI Forensics. *arXiv preprint arXiv:1912.06497*, 2019.

[39] B. Baird, J. Smallwood, M. D. Mrazek, J. W. Kam, M. S. Franklin, and J. W. Schooler. Inspired by distraction: Mind wandering facilitates creative incubation. *Psychological science*, 23(10):1117–1122, 2012.

[40] E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132, 2015.

[41] S. Baldassi, T. Kohno, F. Roesner, and M. Tian. Challenges and new directions in augmented reality, computer security, and neuroscience–part 1: Risks to sensation and perception. *arXiv preprint arXiv:1806.10557*, 2018.

[42] L. D. Ball, G. Ewan, and N. J. Coull. Undermining: social engineering using open source intelligence gathering. In *4th International Conference on Knowledge Discovery and Information Retrieval*, pages 275–280. Scitepress Digital Library, 2012.

[43] P. Barberá and T. Zeitzoff. The new public address system: why do world leaders adopt social media? *International Studies Quarterly*, 62(1):121–130, 2018.

[44] H. Barendregt and A. Raffone. Conscious cognition as a discrete, deterministic, and universal Turing Machine process. 2013.

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

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

[47] J. Barrett, R. Lorenz, and O. Oreshkov. Cyclic quantum causal models. *Nature communications*, 12(1):885, 2021.

[48] L. F. Barrett. Emotions are real. *Emotion*, 12(3):413, 2012.

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

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

[51] 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.

[52] L. F. Barrett. Context reconsidered: Complex signal ensembles, relational meaning, and population thinking in psychological science. *American Psychologist*, 77(8):894, 2022.

[53] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak. Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68, 2019.

[54] L. F. Barrett, K. S. Quigley, and P. Hamilton. An active inference theory of allostasis and interoception in depression. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1708):20160011, 2016.

[55] L. F. Barrett and W. K. Simmons. Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, 16(7):419, 2015.

[56] 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.

[57] L. W. Barsalou. Grounded cognition: Past, present, and future. *Topics in cognitive science*, 2(4):716–724, 2010.

[58] W. Bartley III. The philosophy of Karl Popper. *Philosophia*, 6(3-4):463–494, 1976.

[59] J. Baudrillard. *Simulacra and simulation*. University of Michigan press, 1994.

[60] S. Baum, A. Barrett, and R. V. Yampolskiy. Modeling and interpreting expert disagreement about artificial superintelligence. *Informatica*, 41(7):419–428, 2017.

[61] K. Begley. Beta-testing the ethics plugin. *AI & SOCIETY*, pages 1–3, 2023.

[62] L. Beltran. Quantum Bose–Einstein Statistics for Indistinguishable Concepts in Human Language. *Foundations of Science*, 28(1):43–55, 2023.

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

[64] A. Bessi and E. Ferrara. Social bots distort the 2016 US Presidential election online discussion. *First Monday*, 21(11-7), 2016.

[65] A. Bhat, T. Parr, M. Ramstead, and K. Friston. Immunoceptive inference: why are psychiatric disorders and immune responses intertwined? *Biology & Philosophy*, 36(3):27, 2021.

[66] Y. E. Bigman, D. Wilson, M. N. Arnestad, A. Waytz, and K. Gray. Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General*, 2022.

[67] 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.

[68] D. Blackiston, S. Kriegman, J. Bongard, and M. Levin. Biological Robots: Perspectives on an Emerging Interdisciplinary Field. *arXiv preprint arXiv:2207.00880*, 2022.

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

[70] L. Blackwell, N. Ellison, N. Elliott-Deflo, and R. Schwartz. Harassment in social virtual reality: Challenges for platform governance. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–25, 2019.

[71] H. Blume. Neurodiversity: On the neurological underpinnings of geekdom. *The Atlantic*, 30, 1998.

[72] M. Blythe, E. Encinas, J. Kaye, M. L. Avery, R. McCabe, and K. Andersen. Imaginary design workbooks: Constructive criticism and practical provocation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.

[73] M. Blythe, J. Steane, J. Roe, and C. Oliver. Solutionism, the game: design fictions for positive aging. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3849–3858, 2015.

[74] D. Bohm. *On creativity*. Routledge, 2004.

[75] D. Bohm. *Wholeness and the implicate order*. Routledge, 2005.

[76] 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.

[77] 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.

[78] A. J. Bose and P. Aarabi. Virtual Fakes: DeepFakes for Virtual Reality. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–1. IEEE, 2019.

[79] N. Bostrom. Superintelligence: Paths, dangers, strategies. *Oxford University Press*, 2014.

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

[81] J. Brockman. *Possible minds: Twenty-five ways of looking at AI. Beyond Reward and Punishment. David Deutsch.* Penguin Books, 2020.

[82] 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.

[83] J. Bruineberg, J. Kiverstein, and E. Rietveld. The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, 195(6):2417–2444, 2018.

[84] 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.

[85] C. Bruynsteen, T. Gehring, C. Lupo, J. Bauwelinck, and X. Yin. 100 gbps integrated quantum random number generator based on vacuum fluctuations. *arXiv preprint arXiv:2209.04339*, 2022.

[86] V. Bufacchi. Truth, lies and tweets: a consensus theory of post-truth. *Philosophy & Social Criticism*, page 0191453719896382, 2020.

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

[88] M. Bujić and J. Hamari. Immersive journalism: Extant corpus and future agenda. *CEUR-WS*, 2020.

[89] M. Bujić, M. Salminen, J. Macey, and J. Hamari. "Empathy machine": how virtual reality affects human rights attitudes. *Internet Research*, 2020.

[90] R. C. Bunescu and O. O. Uduehi. Learning to Surprise: A Composer-Audience Architecture. In *ICCC*, pages 41–48, 2019.

[91] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

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

[93] M. Cadoux. AR and VR will make spatial journalism the future of reporting. https://venturebeat.com/2019/11/10/ar-and-vr-will-make-spatial-journalism-the-future-of-reporting/, year = 2019, note = VentureBeat; accessed 04-August-2020.

[94] M. Caldwell, J. Andrews, T. Tanay, and L. Griffin. AI-enabled future crime. *Crime Science*, 9(1):1–13, 2020.

[95] R. A. Calvo and D. Peters. *Positive computing: technology for wellbeing and human potential*. MIT Press, 2014.

[96] D. Cancila, J.-L. Gerstenmayer, H. Espinoza, and R. Passerone. Sharpening the scythe of technological change: Socio-technical challenges of autonomous and adaptive cyber-physical systems. *Designs*, 2(4):52, 2018.

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

[98] N. Caporusso. Deepfakes for the Good: A Beneficial Application of Contentious Artificial Intelligence Technology. In *International Conference on Applied Human Factors and Ergonomics*, pages 235–241. Springer, 2020.

[99] R. L. Carhart-Harris and K. Friston. REBUS and the anarchic brain: toward a unified model of the brain action of psychedelics. *Pharmacological reviews*, 71(3):316–344, 2019.

[100] N. Carlini. A Partial Break of the Honeypots Defense to Catch Adversarial Attacks, 2020.

[101] 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.

[102] N. Carlini and H. Farid. Evading Deepfake-Image Detectors with White-and Black-Box Attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 658–659, 2020.

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

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

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

[106] R. Chapman. Neurodiversity, disability, wellbeing. *Neurodiversity Studies: A New Critical Paradigm*, 2020.

[107] T. Chen, J. Liu, Y. Xiang, W. Niu, E. Tong, and Z. Han. Adversarial attack and defense in reinforcement learning-from ai security view. *Cybersecurity*, 2(1):11, 2019.

[108] X. Chen, J. Liu, H. Zhang, and H. K. Kwan. Cognitive diversity and innovative work behaviour: The mediating roles of task reflexivity and relationship conflict and the moderating role of perceived support. *Journal of Occupational and Organizational Psychology*, 92(3):671–694, 2019.

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

[110] Y. Chen, X. Yuan, J. Zhang, Y. Zhao, S. Zhang, K. Chen, and X. Wang. Devil's Whisper: A General Approach for Physical Adversarial Attacks against Commercial Black-box Speech Recognition Devices. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 2667–2684. USENIX Association, aug 2020.

[111] Y. Cheng, F. Juefei-Xu, Q. Guo, H. Fu, X. Xie, S.-W. Lin, W. Lin, and Y. Liu. Adversarial Exposure Attack on Diabetic Retinopathy Imagery. *arXiv preprint arXiv:2009.09231*, 2020.

[112] E. Cheon and N. M. Su. Futuristic autobiographies: Weaving participant narratives to elicit values around robots. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 388–397, 2018.

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

[114] 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.

[115] S. Chitpin. Should popper's view of rationality be used for promoting teacher knowledge? *Educational Philosophy and Theory*, 45(8):833–844, 2013.

[116] L. Chittaro, R. Sioni, C. Crescentini, and F. Fabbro. Mortality salience in virtual reality experiences and its effects on users' attitudes towards risk. *International Journal of Human-Computer Studies*, 101:10–22, 2017.

[117] A. E. Cinà, A. Torcinovich, and M. Pelillo. A Black-box Adversarial Attack for Poisoning Clustering. *arXiv preprint arXiv:2009.05474*, 2020.

[118] I. Ciosek. AGGRAVATING UNCERTAINTY–RUSSIAN INFORMATION WARFARE IN THE WEST. *Torun International Studies*, 1(13):57–72, 2020.

[119] S. Clarke and J. Whittlestone. A Survey of the Potential Long-term Impacts of AI: How AI Could Lead to Long-term Changes in Science, Cooperation, Power, Epistemics and Values. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 192–202, 2022.

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

[121] S. Cole and E. Maiberg. Deepfake Porn Is Evolving to Give People Total Control Over Women's Bodies. `https://www.vice.com/en/article/9keen8/deepfake-porn-is-evolving-to-give-people-total-control-over-womens-bodies`, 2019. VICE; accessed 08-November-2020.

[122] D. Colins. Disinformation and "Fake News": Interim Report: Government Response to the Committee's Fifth Report of Session 2017–19. *UK House of Commons Digital*, 2018.

[123] E. Colleoni, A. Rozza, and A. Arvidsson. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of communication*, 64(2):317–332, 2014.

[124] 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.

[125] M. Comiter. Attacking artificial intelligence. *Belfer Center Paper*, 8, 2019.

[126] A. Constant, M. J. Ramstead, S. P. Veissière, and K. Friston. Regimes of expectations: An active inference model of social conformity and decision making. *Frontiers in psychology*, 10:679, 2019.

[127] G. E. Corazza. The dynamic universal creativity process. *Dynamic perspectives on creativity: New directions for theory, research, and practice in education*, pages 297–319, 2019.

[128] G. E. Corazza. Beyond the adjacent possible: On the irreducibility of human creativity to biology and physics. *Possibility Studies & Society*, page 27538699221145664, 2023.

[129] G. E. Corazza, S. Agnoli, and S. Mastria. The dynamic creativity framework: Theoretical and empirical investigations. *European Psychologist*, 2022.

[130] G. E. Corazza and T. Lubart. The big bang of originality and effectiveness: A dynamic creativity framework and its application to scientific missions. *Frontiers in Psychology*, 11:575067, 2020.

[131] A. Corcoran, G. Pezzulo, and J. Hohwy. From Allostatic Agents to Counterfactual Cognisers: Active Inference. *Biological Regulation, and The Origins of Cognition. doi*, 10, 2019.

[132] A. W. Corcoran, G. Pezzulo, and J. Hohwy. From allostatic agents to counterfactual cognisers: active inference, biological regulation, and the origins of cognition. *Biology & Philosophy*, 35(3):1–45, 2020.

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

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

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

[136] M. Cortês, S. A. Kauffman, A. R. Liddle, and L. Smolin. The TAP equation: evaluating combinatorial innovation. *arXiv preprint arXiv:2204.14115*, 2022.

[137] K. Crawford, R. Dobbe, T. Dryer, G. Fried, B. Green, E. Kaziunas, A. Kak, V. Mathur, E. McElroy, A. N. Sánchez, et al. AI Now 2019 Report. `https://ainowinstitute.org/AI_Now_2019_Report.pdf`, 2019. AI Now Institute; accessed 23-May-2020.

[138] S. Cresci. A decade of social bot detection. *Communications of the ACM*, 63(10):72–83, 2020.

[139] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th international conference on world wide web companion*, pages 963–972, 2017.

[140] D. H. Cropley, J. C. Kaufman, and A. J. Cropley. Malevolent creativity: A functional model of creativity in terrorism and crime. *Creativity Research Journal*, 20(2):105–115, 2008.

[141] B. Crothers. FBI warns on teenage sextortion as new twists on sex-related scams emerge. `https://www.foxnews.com/tech/fbi-warns-teenage-sextortion-new-twists-sex-scams-emerge`, 2020. Fox News; accessed 02-November-2020.

[142] W. Cui, X. Li, J. Huang, W. Wang, S. Wang, and J. Chen. Substitute model generation for black-box adversarial attack based on knowledge distillation. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 648–652. IEEE, 2020.

[143] T. Cushing. Harrisburg University Researchers Claim Their 'Unbiased' Facial Recognition Software Can Identify Potential Criminals. `https://www.techdirt.com/articles/20200505/17090244442/harrisburg-university-researchers-claim-their-unbiased-facial-recognition-software-can-identify-potential-criminals.shtml`, 2020. techdirt; accessed 02-November-2020.

[144] C. Da Costa. The Women Geniuses Taking on Racial and Gender Bias in AI – and Amazon . `https://www.thedailybeast.com/the-women-geniuses-taking-on-racial-and-gender-bias-in-artificial-intelligence-and-amazon`, 2020. accessed 23-May-2020.

[145] M. H. da Silva, A. C. do Espírito Santo, E. R. Marins, A. P. L. de Siqueira, D. M. Mol, and A. C. de Abreu Mol. Using virtual reality to support the physical security of nuclear facilities. *Progress in Nuclear Energy*, 78:19–24, 2015.

[146] A. Dafoe. AI governance: A research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK*, 2018.

[147] A. M. Dakhel, V. Majdinasab, A. Nikanjam, F. Khomh, M. C. Desmarais, Z. Ming, et al. GitHub Copilot AI pair programmer: Asset or Liability? *arXiv preprint arXiv:2206.15331*, 2022.

[148] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *arXiv preprint arXiv:2011.03395*, 2020.

[149] V. Danry, J. Leong, P. Pataranutaporn, P. Tandon, Y. Liu, R. Shilkrot, P. Punpongsanon, T. Weissman, P. Maes, and M. Sra. AI-Generated Characters: Putting Deepfakes to Good Use. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–5, 2022.

[150] C. de Armas, R. Tori, and A. V. Netto. Use of virtual reality simulators for training programs in the areas of security and defense: a systematic review. *Multimedia Tools and Applications*, 79(5):3495–3515, 2020.

[151] 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.

[152] S. de Haan. Bio-psycho-social interaction: an enactive perspective. *International Review of Psychiatry*, 33(5):471–477, 2021.

[153] N. De la Peña, P. Weil, J. Llobera, E. Giannopoulos, A. Pomés, B. Spanlang, D. Friedman, M. V. Sanchez-Vives, and M. Slater. Immersive journalism: immersive virtual reality for the first-person experience of news. *Presence: Teleoperators and virtual environments*, 19(4):291–301, 2010.

[154] M. DeCamp and C. Lindvall. Latent bias and the implementation of artificial intelligence in medicine. *Journal of the American Medical Informatics Association*, 2020.

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

[156] J. den Houting. Neurodiversity: An insider's perspective, 2019.

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

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

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

[160] 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.

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

[162] A. Dhanda, M. Reina Ortiz, A. Weigert, A. Paladini, A. Min, M. Gyi, S. Su, S. Fai, and M. Santana Quintero. RECREATING CULTURAL HERITAGE ENVIRONMENTS FOR VR USING PHOTOGRAMMETRY. *ISPRS - International Archives*

of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-2/W9:305–310, 2019.

[163] A. Dietrich. *How creativity happens in the brain*. Springer, 2015.

[164] A. Dietrich. Types of creativity. *Psychonomic bulletin & review*, 26(1):1–12, 2019.

[165] A. Dietrich and H. Haider. A neurocognitive framework for human creative thought. *Frontiers in psychology*, 7:2078, 2017.

[166] V. Dignum. AI is multidisciplinary. *AI Matters*, 5(4):18–21, 2020.

[167] R. Duan, X. Ma, Y. Wang, J. Bailey, A. K. Qin, and Y. Yang. Adversarial Camouflage: Hiding Physical-World Attacks with Natural Styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1000–1008, 2020.

[168] G. Dvali, C. Gomez, D. Lüst, Y. Omar, and B. Richter. Universality of black hole quantum computing. *Fortschritte der Physik*, 65(1):1600111, 2017.

[169] G. Dvali and Z. N. Osmanov. Black holes as tools for quantum computing by advanced extraterrestrial civilizations. *arXiv preprint arXiv:2301.09575*, 2023.

[170] M. Eckstein. Conformal cyclic cosmology, gravitational entropy and quantum information. *General Relativity and Gravitation*, 55(2):26, 2023.

[171] K. Epstude and N. J. Roese. The functional theory of counterfactual thinking. *Personality and social psychology review*, 12(2):168–192, 2008.

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

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

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

[175] T. Everitt, G. Lea, and M. Hutter. AGI safety literature review. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 5441–5449, 2018.

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

[177] M. Farella, G. Chiazzese, and G. L. Bosco. Question Answering with BERT: designing a 3D virtual avatar for Cultural Heritage exploration. In *2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON)*, pages 770–774. IEEE, 2022.

[178] H. Farid. Digital forensics in a post-truth age. *Commentary, Forensic Science International*, 289:268–269, 2018.

[179] M. Farokhmanesh. Is It Legal to Swap Someone's Face into Porn without Consent? *Verge. January*, 30, 2018.

[180] T. E. Feinberg and J. Mallatt. The evolutionary and genetic origins of consciousness in the Cambrian Period over 500 million years ago. *Frontiers in psychology*, 4:667, 2013.

[181] R. Feldman. Bio-behavioral synchrony: A model for integrating biological and microsocial behavioral processes in the study of parenting. *Parenting*, 12(2-3):154–164, 2012.

[182] M. Felton. In Search of the Ultimate Model. *Journal of NeuroPhilosophy*, 2(1), 2023.

[183] M. A. Felton Jr. *Universe Within: The Surprising Way the Human Brain Models the Universe*. John Hunt Publishing, 2022.

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

[185] R. Feynman. Cargo cult science: Some remarks on science, pseudoscience, and learning how to not fool yourself-the 1974 Caltech commencement address. *The Best Short Works of Richard P. Feynman-The Pleasure of Finding Things Out*, pages 205–216.

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

[187] T. Field. Relationships as regulators. *Psychology*, 3(06):467, 2012.

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

[189] C. Fields, K. Friston, J. F. Glazebrook, M. Levin, and A. Marcianò. The free energy principle induces neuromorphic development. *Neuromorphic Computing and Engineering*, 2(4):042002, 2022.

[190] A. Fink, M. Benedek, K. Koschutnig, E. Pirker, E. Berger, S. Meister, A. C. Neubauer, I. Papousek, and E. M. Weiss. Training of verbal creativity modulates brain activity in regions associated with language-and memory-related demands. *Human Brain Mapping*, 36(10):4104–4115, 2015.

[191] A. Fink, R. H. Grabner, D. Gebauer, G. Reishofer, K. Koschutnig, and F. Ebner. Enhancing creativity by means of cognitive stimulation: Evidence from an fMRI study. *NeuroImage*, 52(4):1687–1695, 2010.

[192] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019.

[193] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication*, 1(2020), 2020.

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

[195] L. Floridi. Establishing the rules for building trustworthy AI. *Nature Machine Intelligence*, 1(6):261–262, 2019.

[196] 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.

[197] G. Franceschelli and M. Musolesi. Creativity and machine learning: A survey. *arXiv preprint arXiv:2104.02726*, 2021.

[198] V. Franchina and G. L. Coco. The influence of social media use on body image concerns. *International Journal of Psychoanalysis and Education*, 10(1):5–14, 2018.

[199] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz. Leveraging Frequency Analysis for Deep Fake Image Recognition. *arXiv preprint arXiv:2003.08685*, 2020.

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

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

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

[203] S. J. Frenda, E. D. Knowles, W. Saletan, and E. F. Loftus. False memories of fabricated political events. *Journal of Experimental Social Psychology*, 49(2):280–286, 2013.

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

[205] K. Friston. Am I self-conscious?(Or does self-organization entail self-consciousness?). *Frontiers in psychology*, 9:579, 2018.

[206] K. J. Friston, M. Lin, C. D. Frith, G. Pezzulo, J. A. Hobson, and S. Ondobaka. Active inference, curiosity and insight. *Neural computation*, 29(10):2633–2683, 2017.

[207] A. Gardner and J. P. Conlon. Cosmological natural selection and the purpose of the universe. *Complexity*, 18(5):48–56, 2013.

[208] M. Gendron, K. Hoemann, A. N. Crittenden, S. M. Mangola, G. A. Ruark, and L. F. Barrett. Emotion perception in Hadza Hunter-Gatherers. *Scientific reports*, 10(1):1–17, 2020.

[209] A. Giaretta. Security and Privacy in Virtual Reality–A Literature Survey. *arXiv preprint arXiv:2205.00208*, 2022.

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

[211] W. B. Glisson, G. Grispos, and K. K. R. Choo. Cyber Operations, Defence and Forensics. In *54th Annual Hawaii International Conference on System Sciences, HICSS 2021*, pages 6934–6935. IEEE Computer Society, 2021.

[212] J. A. Goldstein, G. Sastry, M. Musser, R. DiResta, M. Gentzel, and K. Sedova. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. *arXiv preprint arXiv:2301.04246*, 2023.

[213] W. Gong and D.-L. Deng. Universal adversarial examples and perturbations for quantum classifiers. *National Science Review*, 9(6):nwab130, 2022.

[214] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[215] K. Goswami, C. Giarmatzi, M. Kewming, F. Costa, C. Branciard, J. Romero, and A. G. White. Indefinite causal order in a quantum switch. *Physical review letters*, 121(9):090503, 2018.

[216] K. Gray, C. Schein, and C. D. Cameron. How to think about emotion and morality: circles, not arrows. *Current opinion in psychology*, 17:41–46, 2017.

[217] K. Gray, A. Waytz, and L. Young. The moral dyad: A fundamental template unifying moral judgment. *Psychological Inquiry*, 23(2):206–215, 2012.

[218] J. Greenberg and J. Arndt. Terror management theory. *Handbook of theories of social psychology*, 1:398–415, 2011.

[219] G. F. Grosu, A. V. Hopp, V. V. Moca, H. Bârzan, A. Ciuparu, M. Ercsey-Ravasz, M. Winkel, H. Linde, and R. C. Mureșan. The fractal brain: scale-invariance in structure and dynamics. *Cerebral Cortex*, 2022.

[220] B. Gu, P. Konana, R. Raghunathan, and H. M. Chen. Research note—The allure of homophily in social media: Evidence from investor responses on virtual communities. *Information Systems Research*, 25(3):604–617, 2014.

[221] M. Guerar, L. Verderame, M. Migliardi, F. Palmieri, and A. Merlo. Gotta CAPTCHA'Em All: A Survey of Twenty years of the Human-or-Computer Dilemma. *arXiv preprint arXiv:2103.01748*, 2021.

[222] A. Gulhane, A. Vyas, R. Mitra, R. Oruche, G. Hoefer, S. Valluripally, P. Calyam, and K. A. Hoque. Security, Privacy and Safety Risk Assessment for Virtual Reality Learning Environment Applications. In *2019 16th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pages 1–9. IEEE, 2019.

[223] E. Gümüş, D. Majidi, D. Nikolić, P. Raif, B. Karimi, J. T. Peltonen, E. Scheer, J. P. Pekola, H. Courtois, W. Belzig, et al. Calorimetry of a phase slip in a josephson junction. *Nature Physics*, pages 1–5, 2023.

[224] E. Gutierrez-Sigut, M. Vergara-Martinez, and M. Perea. The impact of visual cues during visual word recognition in deaf readers: An ERP study. *Cognition*, 218:104938, 2022.

[225] M. Haag and M. Salam. Gunman in 'Pizzagate' Shooting Is Sentenced to 4 Years in Prison. `https://www.nytimes.com/2017/06/22/us/pizzagate-attack-sentence.html`, 2017. The New York Times; accessed 02-November-2017.

[226] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan. Cooperative inverse reinforcement learning. In *Advances in neural information processing systems*, pages 3909–3917, 2016.

[227] T. Hagendorff. The ethics of Ai ethics: An evaluation of guidelines. *Minds and Machines*, pages 1–22, 2020.

[228] A. Halfmann and D. Rieger. Permanently on call: The effects of social pressure on smartphone users' self-control, need satisfaction, and well-being. *Journal of Computer-Mediated Communication*, 24(4):165–181, 2019.

[229] B. Hall. Superintelligence. `http://www.bretthall.org/superintelligence-6.html`, 2020. Part 6: Neologisms and Choices; accessed 04-January-2021.

[230] S. Hameroff. Consciousness, cognition and the neuronal cytoskeleton–A new paradigm needed in neuroscience. *Frontiers in Molecular Neuroscience*, 15, 2022.

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

[232] J. T. Hancock and J. N. Bailenson. The social impact of deepfakes, 2021.

[233] F. Hanusch and D. Nölleke. Journalistic homophily on social media: Exploring journalists' interactions with each other on Twitter. *Digital Journalism*, 7(1):22–44, 2019.

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

[235] K. Hao. The Biggest Threat of Deepfakes Isn't the Deepfakes Themselves, 2019.

[236] K. Hao. A deepfake bot is being used to "undress" underage girls. `https://www.technologyreview.com/2020/10/20/1010789/ai-deepfake-bot-undresses-women-and-underage-girls/`, 2020. MIT Technology Review; accessed 08-November-2020.

[237] J. Happa, M. Glencross, and A. Steed. Cyber security threats and challenges in collaborative mixed-reality. *Frontiers in ICT*, 6:5, 2019.

[238] G. M. Hardee and R. P. McMahan. FIJI: a framework for the immersion-journalism intersection. *Frontiers in ICT*, 4:21, 2017.

[239] S. Harding. *Can theories be refuted?: Essays on the Duhem-Quine thesis*, volume 81. Springer Science & Business Media, 1975.

[240] K. R. Harris. Video on demand: what deepfakes do and how they harm. *Synthese*, 199(5):13373–13391, 2021.

[241] Harrisburg University . HU facial recognition software predicts criminality. `http://archive.is/N1HVe#selection-1509.0-1509.51`, 2020. Online; accessed 23-May-2020.

[242] 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.

[243] 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.

[244] D. Harwell. An artificial-intelligence first: Voice-mimicking software reportedly used in a major theft. `https://www.washingtonpost.com/technology/2019/09/04/an-artificial-intelligence-first-voice-mimicking-software-reportedly-used-major-theft/`, 2019. The Washington Post; accessed 04-August-2020.

[245] S. Z. Hassan, P. Salehi, R. K. Røed, P. Halvorsen, G. A. Baugerud, M. S. Johnson, P. Lison, M. Riegler, M. E. Lamb, C. Griwodz, et al. Towards an AI-driven talking avatar in virtual reality for investigative interviews of children. In *Proceedings of the 2nd Workshop on Games Systems*, pages 9–15, 2022.

[246] J. Hernandez, H. M. Marin-Castro, et al. A Semantic Focused Web Crawler Based on a Knowledge Representation Schema. *Applied Sciences*, 10(11):3837, 2020.

[247] M. Herrero-Collantes and J. C. Garcia-Escartin. Quantum random number generators. *Reviews of Modern Physics*, 89(1):015004, 2017.

[248] 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.

[249] K. Hill. Wrongfully accused by an algorithm. *The New York Times*, 2020.

[250] E. Hine and L. Floridi. New deepfake regulations in China are a tool for social stability, but at what cost? *Nature Machine Intelligence*, 4(7):608–610, 2022.

[251] G. Hinton. The forward-forward algorithm: Some preliminary investigations. *arXiv preprint arXiv:2212.13345*, 2022.

[252] 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.

[253] J. A. Hobson, C. C.-H. Hong, and K. J. Friston. Virtual reality and consciousness inference in dreaming. *Frontiers in psychology*, 5:1133, 2014.

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

[255] 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.

[256] S. M. Hofmann, F. Klotzsche, A. Mariola, V. V. Nikulin, A. Villringer, and M. Gaebler. Decoding subjective emotional arousal from eeg during an immersive virtual reality experience. *bioRxiv*, 2020.

[257] J. Holt-Lunstad. Why social relationships are important for physical health: A systems approach to understanding and modifying risk and protection. *Annual review of psychology*, 69:437–458, 2018.

[258] M. Hoogman, M. Stolte, M. Baas, and E. Kroesbergen. Creativity and ADHD: A review of behavioral studies, the effect of psychostimulants and neural underpinnings. *Neuroscience & Biobehavioral Reviews*, 2020.

[259] 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.

[260] E. Horvitz. On the horizon: Interactive and compositional deepfakes. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 653–661, 2022.

[261] S. Hossenfelder. Maybe the Universe Thinks. Hear Me Out. `https://time.com/6208174/maybe-the-universe-thinks/r`, 2022. TIME; accessed 09-Feburary-2024.

[262] 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.

[263] E. Howell. Humans Really Are Made of Stardust, and a New Study Proves It. *Space News: Science and Astronomy*, 2017.

[264] C. Hu, H.-Q. Xu, and X.-J. Wu. Substitute Meta-Learning for Black-Box Adversarial Attack. *IEEE Signal Processing Letters*, 29:2472–2476, 2022.

[265] B. Huchel. Artificial intelligence examines best ways to keep parolees from recommitting crimes. `https://phys.org/news/2020-08-artificial-intelligence-ways-parolees-recommitting.html`, 2020. Phys Org; accessed 20-August-2020.

[266] J. B. Hutchinson and L. F. Barrett. The power of predictions: An emerging paradigm for psychological research. *Current Directions in Psychological Science*, page 0963721419831992, 2019.

[267] A. Ijjas and P. J. Steinhardt. Entropy, black holes, and the new cyclic universe. *Physics Letters B*, 824:136823, 2022.

[268] M. Illes and P. Wilson. *The scientific method in forensic science: a Canadian Handbook*. Canadian Scholars' Press, 2020.

[269] G. Irving, P. Christiano, and D. Amodei. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

[270] E. Ismagilova, E. Slade, N. P. Rana, and Y. K. Dwivedi. The effect of characteristics of source credibility on consumer behaviour: A meta-analysis. *Journal of Retailing and Consumer Services*, 53, 2020.

[271] D. Izzo, A. Hadjiivanov, D. Dold, G. Meoni, and E. Blazquez. Neuromorphic Computing and Sensing in Space. *arXiv preprint arXiv:2212.05236*, 2022.

[272] N. Jain, A. Olmo, S. Sengupta, L. Manikonda, and S. Kambhampati. Imperfect imaganation: Implications of gans exacerbating biases on facial data augmentation and snapchat selfie lenses. *arXiv preprint arXiv:2001.09528*, 2020.

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

[274] P. Jansen and F. Fischbach. The Social Engineer: An Immersive Virtual Reality Educational Game to Raise Social Engineering Awareness. In *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play*, pages 59–63, 2020.

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

[276] A. John, A. C. Glendenning, A. Marchant, P. Montgomery, A. Stewart, S. Wood, K. Lloyd, and K. Hawton. Self-harm, suicidal behaviours, and cyberbullying in children and young people: Systematic review. *Journal of medical internet research*, 20(4):e129, 2018.

[277] D. G. Johnson. Promises and perils in immersive journalism. *Immersive Journalism as Storytelling: Ethics, Production, and Design*, 2020.

[278] 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.

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

[280] I. Kalpokas and J. Kalpokiene. On alarmism: between infodemic and epistemic anarchy. In *Deepfakes: A Realistic Assessment of Potentials, Risks, and Policy Regulation*, pages 41–53. Springer, 2022.

[281] S. Kang, E. O'Brien, A. Villarreal, W. Lee, and C. Mahood. Immersive Journalism and Telepresence: Does virtual reality news use affect news credibility? *Digital Journalism*, 7(2):294–313, 2019.

[282] I. Kant. Prolegomena to any future metaphysics that may arise as science, 1783.

[283] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[284] A. Kasirer, E. Adi-Japha, and N. Mashal. Verbal and Figural Creativity in Children With Autism Spectrum Disorder and Typical Development. *Frontiers in Psychology*, 11:2968, 2020.

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

[286] H. Katie, K. Zulqarnain, M. J. Feldman, N. Catie, M. Devlin, J. Dy, L. F. Barrett, J. B. Wormwood, and K. S. Quigley. Context-aware experience sampling reveals the scale of variation in affective experience. *Scientific Reports (Nature Publisher Group)*, 10(1), 2020.

[287] S. Kauffman. Eros and logos. In *Ontogenesis Beyond Complexity*, pages 7–21. Routledge, 2021.

[288] S. Kauffman and S. Guerin. Did the Universe Construct Itself? 2023.

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

[290] S. B. Kaufman. Self-Actualizing People in the 21st Century: Integration With Contemporary Theory and Research on Personality and Well-Being. *Journal of Humanistic Psychology*, page 0022167818809187, 2018.

[291] A. Kaushal, R. Altman, and C. Langlotz. Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. *Jama*, 324(12):1212–1213, 2020.

[292] R. Kempsell. Ofqual pauses study into whether AI could be used to mark exams. `https://www.thetimes.co.uk/article/robot-exam-marking-project-is-put-on-hold-vvrm753l3`, 2020. The Times; accessed 10-November-2020.

[293] C. M. Kerskens and D. L. Pérez. Experimental indications of non-classical brain functions. *Journal of Physics Communications*, 6(10):105001, 2022.

[294] O. Keyes. Automating autism: Disability, discourse, and artificial intelligence. *The Journal of Sociotechnical Critique*, 1(1):8, 2020.

[295] C. Kiefer and P. Peter. Time in quantum cosmology. *Universe*, 8(1):36, 2022.

[296] S. Kim, J. Kandampully, and A. Bilgihan. The influence of eWOM communications: An application of online social network framework. *Computers in Human Behavior*, 80:243–254, 2018.

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

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

[299] I. R. Kleckner, J. Zhang, A. Touroutoglou, L. Chanes, C. Xia, W. K. Simmons, K. S. Quigley, B. C. Dickerson, and L. F. Barrett. Evidence for a large-scale brain system supporting allostasis and interoception in humans. *Nature human behaviour*, 1(5):1–14, 2017.

[300] L. Knoepp. Forget Oculus Rift, Meet The Godmother Of VR. `https://www.forbes.com/sites/lillyknoepp/2017/04/13/forget-oculus-rift-meet-the-godmother-of-vr/`, 2017. Forbes; accessed 04-August-2020.

[301] E. Kocabey, F. Ofli, J. Marin, A. Torralba, and I. Weber. Using computer vision to study the effects of BMI on online popularity and weight-based homophily. In *International Conference on Social Informatics*, pages 129–138. Springer, 2018.

[302] M. E. Koltko-Rivera. Rediscovering the later version of Maslow's hierarchy of needs: Self-transcendence and opportunities for theory, research, and unification. *Review of general psychology*, 10(4):302–317, 2006.

[303] Z. Kong, J. Guo, A. Li, and C. Liu. PhysGAN: Generating Physical-World-Resilient Adversarial Examples for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14254–14263, 2020.

[304] M. Koopmans and D. Stamovlasis. Introduction to education as a complex dynamical system. *Complex dynamical systems in education: Concepts, methods and applications*, pages 1–7, 2016.

[305] J. Kotlarek, I.-C. Lin, and K.-L. Ma. Improving spatial orientation in immersive environments. In *Proceedings of the Symposium on Spatial User Interaction*, pages 79–88, 2018.

[306] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown. Text classification algorithms: A survey. *Information*, 10(4):150, 2019.

[307] R. V. Kozinets, A. D. Gershoff, and T. B. White. Introduction to special issue: trust in doubt: consuming in a post-truth world. *Journal of the Association for Consumer Research*, 5(2):130–136, 2020.

[308] 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.

[309] A. Krausová. Czech Republic's AI Observatory and Forum. *The Lawyer Quarterly*, 1(1), 2020.

[310] K. Krishna, G. S. Tomar, A. P. Parikh, N. Papernot, and M. Iyyer. Thieves on sesame street! model extraction of bert-based apis. *arXiv preprint arXiv:1910.12366*, 2019.

[311] E. Krokos, C. Plaisant, and A. Varshney. Virtual memory palaces: immersion aids recall. *Virtual Reality*, 23(1):1–15, 2019.

[312] A. W. Kruglanski, K. Jasko, and K. Friston. All Thinking is 'Wishful' Thinking. *Trends in Cognitive Sciences*, 2020.

[313] M. Kumar, M. Jindal, and M. Kumar. A systematic survey on CAPTCHA recognition: types, creation and breaking techniques. *Archives of Computational Methods in Engineering*, 29(2):1107–1136, 2022.

[314] R. S. S. Kumar, D. O. Brien, K. Albert, S. Viljöen, and J. Snover. Failure modes in machine learning systems. *arXiv preprint arXiv:1911.11034*, 2019.

[315] R. S. S. Kumar, M. Nyström, J. Lambert, A. Marshall, M. Goertzel, A. Comissoneru, M. Swann, and S. Xia. Adversarial Machine Learning–Industry Perspectives. *arXiv preprint arXiv:2002.05646*, 2020.

[316] S. Kurtev. Wave-like patterns in parameter space interpreted as evidence for macroscopic effects resulting from quantum or quantum-like processes in the brain. *Scientific Reports*, 12(1):18938, 2022.

[317] R. Ladhari, E. Massa, and H. Skandrani. YouTube vloggers' popularity and influence: The roles of homophily, emotional attachment, and expertise. *Journal of Retailing and Consumer Services*, 54:102027, 2020.

[318] N. Lahav and Z. A. Neemeh. A relativistic theory of consciousness. *Frontiers in Psychology*, 12:704270, 2022.

[319] C. Laing. A role for onomatopoeia in early language: Evidence from phonological development. *Language and Cognition*, 11(2):173–187, 2019.

[320] A. Lamb. After Covid, AI will Pivot. https://towardsdatascience.com/after-covid-ai-will-pivot-dbe9dd06327, 2020. Towards data sciece; accessed 12-November-2020.

[321] L. Lane. NIST finds flaws in facial checks on people with Covid masks. *Biometric Technology Today*, 2020.

[322] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

[323] S. Lathiya, J. Dhobi, A. Zubiaga, M. Liakata, and R. Procter. Birds of a feather check together: Leveraging homophily for sequential rumour detection. *Online Social Networks and Media*, 19:100097, 2020.

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

[325] M. Lee, P. Liang, and Q. Yang. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. *arXiv preprint arXiv:2201.06796*, 2022.

[326] J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

[327] 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.

[328] J. M. Leonhardt, T. Pezzuti, and J.-E. Namkoong. We're not so different: Collectivism increases perceived homophily, trust, and seeking user-generated product information. *Journal of Business Research*, 112:160–169, 2020.

[329] O. Letter. Our letter to the APA. https://screentimenetwork.org/apa, 2018. Online; accessed 02-November-2020.

[330] Y. Leviathan and Y. Matias. Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone. https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html, 2018.

[331] 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.

[332] J. Li, S. Qu, X. Li, J. Szurley, J. Z. Kolter, and F. Metze. Adversarial music: Real world audio adversary against wake-word detection system. In *Advances in Neural Information Processing Systems*, pages 11931–11941, 2019.

[333] X. Li, Y. Chen, R. Patibanda, and F. Mueller. vrCAPTCHA: exploring CAPTCHA designs in virtual reality. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–4, 2021.

[334] C. Lieber. Tech companies use "persuasive design" to get us hooked. Psychologists say it's unethical. `https://www.vox.com/2018/8/8/17664580/persuasive-technology-psychology`, 2018. Vox; accessed 08-November-2020.

[335] S. Lins, K. D. Pandl, H. Teigeler, S. Thiebes, C. Bayer, and A. Sunyaev. Artificial Intelligence as a Service. *Business & Information Systems Engineering*, 63(4):441–456, 2021.

[336] N. Liv and D. Greenbaum. Deep Fakes and Memory Malleability: False Memories in the Service of Fake News. *AJOB neuroscience*, 11(2):96–104, 2020.

[337] A. Loeb. Endless Creation Out of Nothing – Could our universe have been an experiment by an ancient civilization? . *Scientific American. Oct*, 2020.

[338] A. Loeb. Was Our Universe Created in a Laboratory? *Scientific American. Oct*, 15:2021, 2021.

[339] A. Loeb. Interstellar – The Search for Extraterrestrial Life and Our Future in the Stars., 2023.

[340] E. N. Lorenz. Deterministic nonperiodic flow. *Journal of atmospheric sciences*, 20(2):130–141, 1963.

[341] S. Lu, L.-M. Duan, and D.-L. Deng. Quantum adversarial machine learning. *Physical Review Research*, 2(3):033212, 2020.

[342] D. D. Luxton, J. D. June, and J. M. Fairall. Social media and suicide: a public health perspective. *American journal of public health*, 102(S2):S195–S200, 2012.

[343] R. Mabrook. Collaborative and Experimental Cultures in Virtual Reality Journalism: From the Perspective of Content Creators. *International Journal of Humanities and Social Sciences*, 13(5):532–542, 2019.

[344] T. Macaulay. New AR app will let you model a virtual companion on anyone you want. `https://thenextweb.com/neural/2020/06/01/new-ar-app-will-let-you-model-a-virtual-companion-on-anyone-you-want/`, 2020. Online; accessed 04-August-2020.

[345] M. Madary and T. K. Metzinger. Real virtuality: a code of ethical conduct. Recommendations for good scientific practice and the consumers of VR-technology. *Frontiers in Robotics and AI*, 3:3, 2016.

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

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

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

[349] D. R. Mandel. Chaos theory, sensitive dependence, and the logistic equation. 1995.

[350] C. Marletto. *The Science of Can and Can't: A Physicist's Journey Through the Land of Counterfactuals*. Penguin, 2022.

[351] C. Marletto, V. Vedral, L. T. Knoll, F. Piacentini, E. Bernardi, E. Rebufello, A. Avella, M. Gramegna, I. P. Degiovanni, and M. Genovese. Emergence of constructor-based irreversibility in quantum systems: theory and experiment. *Physical Review Letters*, 128(8):080401, 2022.

[352] S. Martin. AI Observatory opens in Berlin on 3 March 2020 . `https://www.itspmagazine.com/itsp-chronicles/ai-village-what-is-ai-safety-and-how-can-we-embrace-and-prepare-for-adversarial-ai`, 2018. ITSP Magazine; accessed 25-April-2020.

[353] D. Martin Jr, V. Prabhakaran, J. Kuhlberg, A. Smart, and W. S. Isaac. Extending the Machine Learning Abstraction Boundary: A Complex Systems Approach to Incorporate Societal Context. *arXiv preprint arXiv:2006.09663*, 2020.

[354] J. M. M. Martins and M. de Paiva Guimarães. Using Olfactory Stimuli in Virtual Reality Applications. In *2018 20th Symposium on Virtual and Augmented Reality (SVR)*, pages 57–64. IEEE, 2018.

[355] D. McDuff, C. Hurter, and M. Gonzalez-Franco. Pulse and vital sign measurement in mixed reality using a HoloLens. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, pages 1–9, 2017.

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

[357] D. W. Miller. *Out of error: Further essays on critical rationalism*. Ashgate Publishing, Ltd., 2006.

[358] Y. Mirsky and W. Lee. The Creation and Detection of Deepfakes: A Survey. *arXiv preprint arXiv:2004.11138*, 2020.

[359] Y. Mirsky, T. Mahler, I. Shelef, and Y. Elovici. CT-GAN: Malicious tampering of 3D medical imagery using deep learning. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 461–478, 2019.

[360] C. W. Misner, K. S. Thorne, and W. H. Zurek. John Wheeler, relativity, and quantum information. *Physics Today*, 62(4):40–46, 2009.

[361] 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.

[362] R. Mitchell, B. Boyle, R. O'Brien, A. Malik, K. Tian, V. Parker, M. Giles, P. Joyce, and V. Chiang. Balancing cognitive diversity and mutual understanding in multidisciplinary teams. *Health care management review*, 42(1):42–52, 2017.

[363] B. Mittelstadt. AI Ethics–Too principled to fail. *arXiv preprint arXiv:1906.06668*, 2019.

[364] J. Mossbridge, B. Goertzel, R. Mayet, E. Monroe, G. Nehat, D. Hanson, and G. Yu. Emotionally-sensitive AI-driven android interactions improve social welfare through helping people access self-transcendent states. vol. In *AI for Social Good Workshop at Neural Information Processing Systems 2018 Conference*, 2018.

[365] P. Mozur. China's 'hybrid war': Beijing's mass surveillance of Australia and the world for secrets and scandal. `https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html`, 2019. The New York Times; accessed 04-August-2020.

[366] I. Q. Mundial, M. S. U. Hassan, M. I. Tiwana, W. S. Qureshi, and E. Alanazi. Towards Facial Recognition Problem in COVID-19 Pandemic. In *2020 4rd International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM)*, pages 210–214. IEEE, 2020.

[367] G. Murphy, E. F. Loftus, R. H. Grady, L. J. Levine, and C. M. Greene. False memories for fake news during Ireland's abortion referendum. *Psychological science*, 30(10):1449–1459, 2019.

[368] T. Nakamoto, T. Hirasawa, and Y. Hanyu. Virtual environment with smell using wearable olfactory display and computational fluid dynamics simulation. In *2020*

IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pages 713–720. IEEE, 2020.

[369] B. Nassi, D. Nassi, R. Ben-Netanel, Y. Mirsky, O. Drokin, and Y. Elovici. Phantom of the ADAS: Phantom Attacks on Driver-Assistance Systems. *IACR Cryptol. ePrint Arch.*, 2020:85, 2020.

[370] P. Neekhara, S. Hussain, M. Jere, F. Koushanfar, and J. McAuley. Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples. *arXiv preprint arXiv:2002.12749*, 2020.

[371] S. D. Nelson, Esq., and J. W. Simek. Video and Audio Deepfakes: What Lawyers Need to Know. `https://www.masslomap.org/video-and-audio-deepfakes-what-lawyers-need-to-know-guest-post/`, 2020. Sensei Enterprises, Inc.; accessed 08-November-2020.

[372] M. L. Ngan, P. J. Grother, and K. K. Hanaoka. Ongoing Face Recognition Vendor Test (FRVT) Part 6A: Face recognition accuracy with masks using pre-COVID-19 algorithms. *National Institute of Standards and Technology*, 2020.

[373] S. Niedenthal, P. Lundén, M. Ehrndal, and J. K. Olofsson. A handheld olfactory display for smell-enabled VR games. In *2019 IEEE International Symposium on Olfaction and Electronic Nose (ISOEN)*, pages 1–4. IEEE, 2019.

[374] M. G. Nilsson, K. T. Pepelasi, M. Ioannou, and D. Lester. Understanding the link between Sextortion and Suicide. *International Journal of Cyber Criminology*, 13(1):55–69, 2019.

[375] D. Noble. The role of stochasticity in biological communication processes. *Progress in Biophysics and Molecular Biology*, 162:122–128, 2021.

[376] R. Noble and D. Noble. Harnessing stochasticity: How do organisms make choices? *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(10):106309, 2018.

[377] R. Noble and D. Noble. Can reasons and values influence action: how might intentional agency work physiologically? *Journal for General Philosophy of Science*, 52:277–295, 2021.

[378] J. Norman and Y. Bar-Yam. Special Operations Forces: A Global Immune System? In *International Conference on Complex Systems*, pages 486–498. Springer, 2018.

[379] A. North Whitehead. Process and reality: an essay in Cosmology, 1929.

[380] B. A. Nosek and D. Lakens. Registered reports, 2014.

[381] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[382] M. Obradović, I. Vasiljević, I. DJurić, J. Kićanović, V. Stojaković, and R. Obradović. Virtual Reality Models Based on Photogrammetric Surveys – A Case Study of the Iconostasis of the Serbian Orthodox Cathedral Church of Saint Nicholas in Sremski Karlovci (Serbia). *Applied Sciences*, 10(8):2743, 2020.

[383] F. O'Brolcháin, T. Jacquemard, D. Monaghan, N. O'Connor, P. Novitzky, and B. Gordijn. The convergence of virtual reality and social networks: threats to privacy and autonomy. *Science and engineering ethics*, 22(1):1–29, 2016.

[384] L. O'Donnell. Black Hat 2020: Open-Source AI to Spur Wave of 'Synthetic Media' Attacks. `https://threatpost.com/black-hat-2020-open-source-ai-to-spur-wave-of-synthetic-media-attacks/158066/`, 2020. threatpost; accessed 08-November-2020.

[385] OECD.AI. OECD AI Policy Observatory . `https://oecd.ai/`, 2020. Online; accessed 25-April-2020.

[386] 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.

[387] C. Öhman. Introducing the pervert's dilemma: a contribution to the critique of Deepfake Pornography. *Ethics and Information Technology*, pages 1–8, 2019.

[388] B. Ohst and B. Tuschen-Caffier. Does physiological arousal lead to increased catastrophic misinterpretation? An experiment based on the concept of a fear memory. *BMC psychology*, 8(1):1–11, 2020.

[389] S. Oosterwijk, K. A. Lindquist, E. Anderson, R. Dautoff, Y. Moriguchi, and L. F. Barrett. States of mind: Emotions, body feelings, and thoughts share distributed neural networks. *NeuroImage*, 62(3):2110–2128, 2012.

[390] 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.

[391] E. J. Oughton, D. Ralph, R. Pant, E. Leverett, J. Copic, S. Thacker, R. Dada, S. Ruffle, M. Tuveson, and J. W. Hall. Stochastic Counterfactual Risk Analysis for the Vulnerability Assessment of Cyber-Physical Attacks on Electricity Distribution Infrastructure Networks. *Risk Analysis*, 39(9):2012–2031, 2019.

[392] 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.

[393] P. K. Palit. Swami Vivekananda. In *Reappraising Modern Indian Thought: Themes and Thinkers*, pages 73–100. Springer, 2022.

[394] T. N. Palmer. Invariant set theory. *arXiv preprint arXiv:1605.01051*, 2016.

[395] R. V. Palumbo, M. E. Marraccini, L. L. Weyandt, O. Wilder-Smith, H. A. McGee, S. Liu, and M. S. Goodwin. Interpersonal autonomic physiology: A systematic review of the literature. *Personality and Social Psychology Review*, 21(2):99–141, 2017.

[396] K. A. Pantserev. The Malicious Use of AI-Based Deepfake Technology as the New Threat to Psychological Security and Political Stability. In *Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity*, pages 37–55. Springer, 2020.

[397] P. Paola, G. Laura, M. Giusy, and C. Michela. Autism, autistic traits and creativity: a systematic review and meta-analysis. *Cognitive Processing*, pages 1–36, 2020.

[398] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman. Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814*, 2016.

[399] H. Päs. *The One: How an Ancient Idea Holds the Future of Physics*. New York, NY: Basic Books, 2023.

[400] L. Pascu. Biometric software that allegedly predicts criminals based on their face sparks industry controversy. `https://www.biometricupdate.com/202005/biometric-software-that-allegedly-predicts-criminals-based-on-their-face-sparks-industry-controversy`, 2020. Biometric; accessed 23-May-2020.

[401] J. V. Pavlik. *Journalism in the age of virtual reality: How experiential media are transforming news*. Columbia University Press, 2019.

[402] J. V. Pavlik. Drones, augmented reality and virtual reality journalism: Mapping their role in immersive news content. *Media and Communication*, 8(3):137–146, 2020.

[403] H. Pearce, B. Ahmad, B. Tan, B. Dolan-Gavitt, and R. Karri. Asleep at the keyboard? assessing the security of github copilot's code contributions. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 754–768. IEEE, 2022.

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

[405] D. Peters, K. Vold, D. Robinson, and R. A. Calvo. Responsible AI—Two Frameworks for Ethical Design Practice. *IEEE Transactions on Technology and Society*, 1(1):34–47, 2020.

[406] A. G. Pillai, N. Ahmadpour, S. Yoo, A. B. Kocaballi, S. Pedell, V. P. Sermuga Pandian, and S. Suleri. Communicate, Critique and Co-create (CCC) Future Technologies through Design Fictions in VR Environment. In *Companion Publication of the 2020 ACM Designing Interactive Systems Conference*, pages 413–416, 2020.

[407] F. Pistono and R. V. Yampolskiy. Unethical Research: How to Create a Malevolent Artificial Intelligence. *CoRR*, abs/1605.02817, 2016.

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

[409] G. Polya. Fake news: "fake realities" and lying by omission. *Global Research*, 18, 2018.

[410] K. Popper. *The Poverty of Historicism*. Routledge & Kegan Paul Ltd., London, 1957.

[411] K. Popper. *Conjectures and refutations: The growth of scientific knowledge*. Routledge, 1963.

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

[413] S. Povolny and S. Trivedi. Model Hacking ADAS to Pave Safer Roads for Autonomous Vehicles. `https://www.mcafee.com/blogs/other-blogs/mcafee-labs/model-hacking-adas-to-pave-safer-roads-for-autonomous-vehicles/`, 2020. McAfee; accessed 08-November-2020.

[414] V. U. Prabhu and A. Birhane. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*, 2020.

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

[416] A. Probyn and M. Doran. One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority. `https://www.abc.net.au/news/2020-09-14/chinese-data-leak-linked-to-military-names-australians/12656668`, 2020. ABC News; accessed 04-August-2020.

[417] H. Qiu, C. Xiao, L. Yang, X. Yan, H. Lee, and B. Li. Semanticadv: Generating adversarial examples via attribute-conditional image editing. *arXiv preprint arXiv:1906.07927*, 2019.

[418] QUARTZ. Virtual reality, fake news and the future of fact. https://www.youtube.com/watch?v=i5LW03vw_x8, 2017. YouTube video; accessed 04-August-2020.

[419] 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.

[420] A. Rahman, M. S. Hossain, N. A. Alrajeh, and F. Alsolami. Adversarial examples–security threats to COVID-19 deep learning systems in medical IoT devices. *IEEE Internet of Things Journal*, 2020.

[421] J. A. Raines and D. E. Litt. Olfactory simulation system for head-mounted displays, Apr. 30 2020. US Patent App. 16/670,572.

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

[423] J. Rajendran, V. Jyothi, and R. Karri. Blue team red team approach to hardware trust assessment. In *2011 IEEE 29th international conference on computer design (ICCD)*, pages 285–288. IEEE, 2011.

[424] 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.

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

[426] A. Rapp. Design fictions for learning: A method for supporting students in reflecting on technology in Human-Computer Interaction courses. *Computers & Education*, 145:103725, 2020.

[427] S. Ray. The Cell: A molecular approach. *The Yale journal of biology and medicine*, 87(4):603, 2014.

[428] A. Rege. Incorporating the human element in anticipatory and dynamic cyber defense. In *2016 IEEE International Conference on Cybercrime and Computer Forensic (ICCCF)*, pages 1–7. IEEE, 2016.

[429] W. Ren, W. Li, S. Xu, K. Wang, W. Jiang, F. Jin, X. Zhu, J. Chen, Z. Song, P. Zhang, et al. Experimental quantum adversarial learning with programmable superconducting qubits. *Nature Computational Science*, 2(11):711–717, 2022.

[430] J. Renshon, J. J. Lee, and D. Tingley. Physiological arousal and political beliefs. *Political Psychology*, 36(5):569–585, 2015.

[431] Reuters. Deepfake Used to Attack Activist Couple Shows New Disinformation Frontier. https://gadgets.ndtv.com/internet/features/deepfake-oliver-taylor-mazen-masri-terrorist-accuse-london-university-of-birmingham-student-fake-profile-22640449, 2020. Reuters; accessed 08-November-2020.

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

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

[434] A. Riikonen. Decide, Disrupt, Destroy: Information Systems in Great Power Competition with China. *Strategic Studies Quarterly*, 13(4), 2019.

[435] H. L. Roediger, F. M. Zaromb, and W. Lin. 1.02 - A Typology of Memory Terms. In J. H. Byrne, editor, *Learning and Memory: A Comprehensive Reference (Second Edition)*, pages 7 – 19. Academic Press, Oxford, 2 edition, 2017.

[436] N. J. Roese and K. Epstude. The functional theory of counterfactual thinking: New evidence, new challenges, new insights. In *Advances in experimental social psychology*, volume 56, pages 1–79. Elsevier, 2017.

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

[438] A. Roli, J. Jaeger, and S. A. Kauffman. How organisms come to know the world: fundamental limits on artificial general intelligence. *Frontiers in Ecology and Evolution*, 9:1035, 2022.

[439] A. A. Rosenberg, M. Halpern, S. Shulman, C. Wexler, and P. Phartiyal. Reinvigorating the role of science in democracy. *PLoS Biol*, 11(5):e1001553, 2013.

[440] D. Rudrauf, D. Bennequin, I. Granic, G. Landini, K. Friston, and K. Williford. A mathematical model of embodied consciousness. *Journal of theoretical biology*, 428:106–131, 2017.

[441] D. Rudrauf, D. Bennequin, and K. Williford. The Moon Illusion explained by the Projective Consciousness Model. *Journal of Theoretical Biology*, page 110455, 2020.

[442] E. Rushing. A Philly lawyer nearly wired $9,000 to a stranger impersonating his son's voice, showing just how smart scammers are getting. https://www.inquirer.com/news/voice-scam-impersonation-fraud-

bail-bond-artificial-intelligence-20200309.html, 2020. The Philadelphia Inquirer; accessed 04-August-2020.

[443] O. S. Torus interconnect. https://en.wikipedia.org/wiki/Torus_interconnect, 2022. Wikipedia; accessed 22-February-2022.

[444] S. Sadaghiani and T. H. Alderson. Tangling with the entangled brain: Putting the global back into the local. *Journal of Cognitive Neuroscience*, 35(3):365–367, 2023.

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

[446] N. Sajid, P. J. Ball, and K. J. Friston. Active inference: demystified and compared. *arXiv*, pages arXiv–1909, 2019.

[447] M. J. Saks and J. J. Koehler. The individualization fallacy in forensic science evidence. *Vand. L. Rev.*, 61:199, 2008.

[448] A. Saleem and A. Ellahi. Influence of electronic word of mouth on purchase intention of fashion products in social networking websites. *Pakistan Journal of Commerce and Social Sciences (PJCSS)*, 11(2):597–622, 2017.

[449] A. L. Sánchez Laws. Can immersive journalism enhance empathy? *Digital Journalism*, 8(2):213–228, 2020.

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

[451] R. Satter. Experts: Spy used AI-generated face to connect with targets. https://apnews.com/bc2f19097a4c4fffaa00de6770b8a60d, 2019. Associated Press (AP); accessed 04-August-2020.

[452] R. Satter. Deepfake Used to Attack Activist Couple Shows New Disinformation Frontier. https://www.reuters.com/article/us-cyber-deepfake-activist/deepfake-used-to-attack-activist-couple-shows-new-disinformation-frontier-idUSKCN24G15E, 2020. Reuters; accessed 04-August-2020.

[453] 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.

[454] D. E. Saxbe, L. Beckes, S. A. Stoycos, and J. A. Coan. Social Allostasis and Social Allostatic Load: A New Model for Research in Social Dynamics, Stress, and Health. *Perspectives on Psychological Science*, 15(2):469–482, 2020.

[455] D. A. Sbarra and C. Hazan. Coregulation, dysregulation, self-regulation: An integrative analysis and empirical agenda for understanding adult attachment, separation, loss, and recovery. *Personality and Social Psychology Review*, 12(2):141–167, 2008.

[456] 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.

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

[458] J. Schneider and F. Breitinger. AI Forensics: Did the Artificial Intelligence System Do It? Why? *arXiv preprint arXiv:2005.13635*, 2020.

[459] W. T. Schneider, R. A. Holland, and O. Lindecke. Over 50 years of behavioural evidence on the magnetic sense in animals: what has been learnt? *The European Physical Journal Special Topics*, pages 1–10, 2023.

[460] E. Schrödinger. *My view of the world*. Cambridge University Press, 2008.

[461] P. J. Scott and R. V. Yampolskiy. Classification Schemas for Artificial Intelligence Failures. *Delphi-Interdisciplinary Review of Emerging Technologies*, 2(4):186–199, 2020.

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

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

[464] K. Y. Segovia and J. N. Bailenson. Virtually true: Children's acquisition of false memories in virtual reality. *Media Psychology*, 12(4):371–393, 2009.

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

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

[467] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer. The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9, 2018.

[468] O. Shehryar and D. M. Hunt. A terror management perspective on the persuasiveness of fear appeals. *Journal of consumer psychology*, 15(4):275–287, 2005.

[469] I. Shumailov, Y. Zhao, D. Bates, N. Papernot, R. Mullins, and R. Anderson. Sponge Examples: Energy-Latency Attacks on Neural Networks. *arXiv preprint arXiv:2006.03463*, 2020.

[470] H. T. Siegelmann. Complex systems science and brain dynamics. *Frontiers in Computational Neuroscience*, 4:7, 2010.

[471] D. K. Simonton. Creative thought as blind variation and selective retention: Why creativity is inversely related to sightedness. *Journal of Theoretical and Philosophical Psychology*, 33(4):253, 2013.

[472] N. K. Singh, D. S. Tomar, and A. K. Sangaiah. Sentiment analysis: a review and comparative analysis over social media. *Journal of Ambient Intelligence and Humanized Computing*, 11(1):97–117, 2020.

[473] M. Slater, C. Gonzalez-Liencres, P. Haggard, C. Vinkers, R. Gregory-Clarke, S. Jelley, Z. Watson, G. Breen, R. Schwarz, W. Steptoe, et al. The ethics of realism in virtual and augmented reality. *Frontiers in Virtual Reality*, 1:1, 2020.

[474] L. Slocombe, M. Sacchi, and J. Al-Khalili. An open quantum systems approach to proton tunnelling in DNA. *Communications Physics*, 5(1):109, 2022.

[475] G. Smith and I. Rustagi. The Problem With COVID-19 Artificial Intelligence Solutions and How to Fix Them. `https://ssir.org/articles/entry/the_problem_with_covid_19_artificial_intelligence_solutions_and_how_to_fix_them`, 2020. Standford Social Innovation Review; accessed 12-November-2020.

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

[477] L. Smolin. Cosmological natural selection as the explanation for the complexity of the universe. *Physica A: Statistical Mechanics and Its Applications*, 340(4):705–713, 2004.

[478] N. Soares and B. Fallenstein. Agent foundations for aligning machine intelligence with human interests: a technical research agenda. In *The Technological Singularity*, pages 103–125. Springer, 2017.

[479] S. Solomon, J. Greenberg, and T. Pyszczynski. A terror management theory of social behavior: The psychological functions of self-esteem and cultural worldviews. In *Advances in experimental social psychology*, volume 24, pages 93–159. Elsevier, 1991.

[480] S. Solomon, J. Greenberg, and T. Pyszczynski. *The worm at the core: On the role of death in life*. Random House, 2015.

[481] N. Spatola and K. Urbanska. God-like robots: the semantic overlap between representation of divine and artificial entities. *Ai & Society*, 35:329–341, 2020.

[482] G. Spocchia. Republican candidate shares conspiracy theory that George Floyd murder was faked. `https://www.independent.co.uk/news/world/americas/us-politics/george-floyd-murder-fake-conspiracy-theory-hoax-republican-gop-missouri-a9580896.html`, 2020. Independent; accessed 04-August-2020.

[483] S. Srivastava. Analysing the Futuristic Potentials of Deepfake + Augmented and Virtual Reality. `https://www.analyticsinsight.net/analysing-the-futuristic-potentials-of-deepfake-augmented-and-virtual-reality/`, 2020. Analytics Insight; accessed 04-August-2020.

[484] M. Stepanovic and V. Ferraro. Reflecting on New Approaches for the Design for Behavioural Change Research and Practice: Shaping the Technologies Through Immersive Design Fiction Prototyping. In *International Conference on Human-Computer Interaction*, pages 542–560. Springer, 2020.

[485] S. Stieger and D. Lewetz. A week without using social media: Results from an ecological momentary intervention study using smartphones. *Cyberpsychology, Behavior, and Social Networking*, 21(10):618–624, 2018.

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

[487] E. Sullivan, M. Sondag, I. Rutter, W. Meulemans, S. Cunningham, B. Speckmann, and M. Alfano. Vulnerability in social epistemic networks. *International Journal of Philosophical Studies*, 28(5):731–753, 2020.

[488] R. M. Sullivan, D. A. Wilson, N. Ravel, and A.-M. Mouly. Olfactory memory networks: from emotional learning to social behaviors. *Frontiers in behavioral neuroscience*, 9:36, 2015.

[489] Q. Sun, A. Tewari, W. Xu, M. Fritz, C. Theobalt, and B. Schiele. A hybrid model for identity obfuscation by face replacement. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 553–569, 2018.

[490] 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.

[491] J. Taylor. Facebook incorrectly removes picture of Aboriginal men in chains because of 'nudity' . https://www.theguardian.com/technology/2020/jun/13/facebook-incorrectly-removes-picture-of-aboriginal-men-in-chains-because-of-nudity, 2020. The Guardian; accessed 02-November-2020.

[492] C. T. Thanh and I. Zelinka. A Survey on Artificial Intelligence in Malware as Next-Generation Threats. In *Mendel*, volume 25, pages 27–34, 2019.

[493] The Agency for Digital Italy. Italian Observatory on Artificial Intelligence . https://ia.italia.it/en/ai-observatory/, 2020. Online; accessed 25-April-2020.

[494] J. E. Theriault, L. Young, and L. F. Barrett. The sense of should: A biologically-based framework for modeling social pressure. *Physics of Life Reviews*, 2020.

[495] 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.

[496] A. Tomar and S. K. Malik. *Reappraising Modern Indian Thought: Themes and Thinkers*. Springer, 2022.

[497] F. Tramer, N. Carlini, W. Brendel, and A. Madry. On adaptive attacks to adversarial example defenses. *arXiv preprint arXiv:2002.08347*, 2020.

[498] M. Trazzi and R. V. Yampolskiy. Building safer AGI by introducing artificial stupidity. *arXiv preprint arXiv:1808.03644*, 2018.

[499] M. Trazzi and R. V. Yampolskiy. Artificial Stupidity: Data We Need to Make Machines Our Equals. *Patterns*, 1(2):100021, 2020.

[500] I. Tribusean. The Use of VR in Journalism: Current Research and Future Opportunities. In *Augmented Reality and Virtual Reality*, pages 227–239. Springer, 2020.

[501] J. Tsao, C. Ting, and C. Johnson. Creative outcome as implausible utility. *Review of General Psychology*, 23(3):279–292, 2019.

[502] G. Tsaramirsis, M. Papoutsidakis, M. Derbali, F. Q. Khan, and F. Michailidis. Towards Smart Gaming Olfactory Displays. *Sensors*, 20(4):1002, 2020.

[503] W.-J. Tseng, E. Bonnail, M. Mcgill, M. Khamis, E. Lecolinet, S. Huron, and J. Gugenheimer. The Dark Side of Perceptual Manipulations in Virtual Reality. In *CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2022.

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

[505] P. Tully and L. Foster. Repurposing Neural Networks to Generate Synthetic Media for Information Operations. https://www.blackhat.com/us-20/briefings/schedule/#repurposing-neural-networks-to-generate-synthetic-media-for-information-operations-19529, 2020. Session at blackhat USA 2020; accessed 08-August-2020.

[506] A. Turchin, D. Denkenberger, and B. P. Green. Global Solutions vs. Local Solutions for the AI Safety Problem. *Big Data and Cognitive Computing*, 3(1):16, 2019.

[507] A. M. Turing and J. Haugeland. *Computing machinery and intelligence*. MIT Press Cambridge, MA, 1950.

[508] J. Uesato, B. O'donoghue, P. Kohli, and A. Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *International Conference on Machine Learning*, pages 5025–5034. PMLR, 2018.

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

[510] T. Uskali, A. Gynnild, S. Jones, and E. Sirkkunen. *Immersive Journalism as Storytelling: Ethics, Production, and Design*. Routledge, 2020.

[511] T. Uskali and P. Ikonen. THE IMPACT OF EMOTIONS IN IMMERSIVE JOURNALISM. *Immersive Journalism as Storytelling: Ethics, Production, and Design*, 2020.

[512] UW Allen School Security and Privacy Research Lab. 2019 Industry-Academia Summit on Mixed Reality Security, Privacy, and Safety: Summit Report . https://ar-sec.cs.washington.edu/research.html, 2019. Online; accessed 04-August-2020.

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

[514] V. Vanchurin. The world as a neural network. *Entropy*, 22(11):1210, 2020.

[515] F. Vazza and A. Feletti. The quantitative comparison between the neuronal network and the cosmic web. *Frontiers in Physics*, 8:525731, 2020.

[516] A. E. Venema and Z. J. Geradts. Digital Forensics, Deepfakes, and the Legal Process. *TheSciTechLawyer*, 16(4):14–23, 2020.

[517] S. V. Veneruso, L. S. Ferro, A. Marrella, M. Mecella, and T. Catarci. CyberVR: an interactive learning experience in virtual reality for cybersecurity related issues. In *Proceedings of the International Conference on Advanced Visual Interfaces*, pages 1–8, 2020.

[518] J. L. Vernon. Understanding the butterfly effect. *American Scientist*, 105(3):130, 2017.

[519] 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.

[520] P. Virtue, X. Y. Stella, and M. Lustig. Better than real: Complex-valued neural nets for MRI fingerprinting. In *2017 IEEE international conference on image processing (ICIP)*, pages 3953–3957. IEEE, 2017.

[521] V. Visoottiviseth, A. Phungphat, N. Puttawong, P. Chantaraumporn, and J. Haga. Lord of secure: the virtual reality game for educating network security. In *2018 seventh ict international student project conference (ict-ispc)*, pages 1–6. IEEE, 2018.

[522] S. Vivekananda. The complete works of Swami Vivekananda, vol. I–VIII. *Advaita Ashrama, Calcutta*, 1907.

[523] R. Vonk. Effects of stereotypes on attitude inference: Outgroups are black and white, ingroups are shaded. *British Journal of Social Psychology*, 41(1):157–167, 2002.

[524] 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.

[525] M. M. Waldrop. *Complexity: The emerging science at the edge of order and chaos*. Simon and Schuster, 1993.

[526] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh. Universal adversarial triggers for NLP. *arXiv preprint arXiv:1908.07125*, 2019.

[527] M. Wang, X.-Q. Lyu, Y.-J. Li, and F.-L. Zhang. VR content creation and exploration with deep learning: A survey. *Computational Visual Media*, pages 1–26, 2020.

[528] Y. Wang, J. Amores, and P. Maes. On-Face Olfactory Interfaces. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2020.

[529] Y. Wang, H. Lv, X. Kuang, G. Zhao, Y.-a. Tan, Q. Zhang, and J. Hu. Towards a Physical-World Adversarial Patch for Blinding Object Detection Models. *Information Sciences*, 2020.

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

[531] G. Weidman. *Penetration testing: a hands-on introduction to hacking*. No Starch Press, 2014.

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

[533] J. D. West and C. T. Bergstrom. Misinformation in and about science. *Proceedings of the National Academy of Sciences*, 118(15):e1912444117, 2021.

[534] H. A. White. Thinking "Outside the Box": Unconstrained Creative Generation in Adults with Attention Deficit Hyperactivity Disorder. *The Journal of Creative Behavior*, 54(2):472–483, 2020.

[535] H. A. White and P. Shah. Scope of semantic activation and innovative thinking in college students with ADHD. *Creativity Research Journal*, 28(3):275–282, 2016.

[536] J. Whittlestone, R. Nyrup, A. Alexandrova, and S. Cave. The role and limits of principles in AI ethics: towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 195–200, 2019.

[537] C. Whyte. Deepfake news: AI-enabled disinformation as a multi-level public policy challenge. *Journal of Cyber Policy*, 5(2):199–217, 2020.

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

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

[540] J. B. Wormwood, E. H. Siegel, J. Kopec, K. S. Quigley, and L. F. Barrett. You are what I feel: A test of the affective realism hypothesis. *Emotion*, 19(5):788, 2019.

[541] J. Wu, M. Zhou, S. Liu, Y. Liu, and C. Zhu. Decision-based Universal Adversarial Attack. *arXiv preprint arXiv:2009.07024*, 2020.

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

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

[544] S. Xu and A. Zhou. Hashtag homophily in twitter network: Examining a controversial cause-related marketing campaign. *Computers in Human Behavior*, 102:87–96, 2020.

[545] K. C. Yam, Y. E. Bigman, P. M. Tang, R. Ilies, D. De Cremer, H. Soh, and K. Gray. Robots at work: People prefer—and forgive—service robots with perceived feelings. *Journal of Applied Psychology*, 2020.

[546] R. Yampolskiy. Usable Guidelines Aim to Make AI Safer. `https://www.mouser.com/blog/usable-guidelines-aim-to-make-ai-safer`, 2020. All, EIT 2020: The Intelligent Revolution; accessed 13-November-2020.

[547] R. V. Yampolskiy. Analyzing user password selection behavior for reduction of password space. In *Proceedings 40th Annual 2006 International Carnahan Conference on Security Technology*, pages 109–115. IEEE, 2006.

[548] R. V. Yampolskiy. Mimicry attack on strategy-based behavioral biometric. In *Fifth International Conference on Information Technology: New Generations (itng 2008)*, pages 916–921. IEEE, 2008.

[549] R. V. Yampolskiy. Taxonomy of pathways to dangerous artificial intelligence. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[550] R. V. Yampolskiy. Predicting future AI failures from historic examples. *foresight*, 2019.

[551] R. V. Yampolskiy. Unpredictability of AI. *arXiv preprint arXiv:1905.13053*, 2019.

[552] R. V. Yampolskiy. On Controllability of AI. *arXiv preprint arXiv:2008.04071*, 2020.

[553] R. V. Yampolskiy, L. Ashby, and L. Hassan. Wisdom of Artificial Crowds—A Metaheuristic Algorithm for Optimization. *Journal of Intelligent Learning Systems and Applications*, 4:98–107, 2012.

[554] R. V. Yampolskiy and V. Govindaraju. Taxonomy of behavioural biometrics. In *Behavioral Biometrics for Human Identification: Intelligent Applications*, pages 1–43. IGI Global, 2010.

[555] R. V. Yampolskiy and M. Spellchecker. Artificial intelligence safety and cybersecurity: A timeline of AI failures. *arXiv preprint arXiv:1610.07997*, 2016.

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

[557] D. Yang, Y. Zhou, Z. Zhang, T. J.-J. Li, and R. LC. AI as an Active Writer: Interaction strategies with generated text in human-AI collaborative fiction writing. In *Joint Proceedings of the ACM IUI Workshops 2022*, volume 10, 2022.

[558] K.-C. Yang, O. Varol, C. A. Davis, E. Ferrara, A. Flammini, and F. Menczer. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1(1):48–61, 2019.

[559] J. Yoo. Ideological Homophily and Echo Chamber Effect in Internet and Social Media. *Student International Journal of Research*, 4(1):1–7, 2007.

[560] L. Young. Calibration Camouflage: Hyphen-Labs and Adam Harvey: HyperFace. *Architectural Design*, 89(1):28–31, 2019.

[561] A. Yuan, A. Coenen, E. Reif, and D. Ippolito. Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces*, pages 841–852, 2022.

[562] H. Zadeh-Haghighi and C. Simon. Magnetic field effects in biology from the perspective of the radical pair mechanism. *Journal of the Royal Society Interface*, 19(193):20220325, 2022.

[563] S. Zafari and S. T. Koeszegi. Attitudes toward attributed agency: Role of perceived control. *International Journal of Social Robotics*, 13(8):2071–2080, 2021.

[564] J. Zang, L. Sweeney, and M. Weiss. The real threat of fake voices in a time of crisis. `https://techcrunch.com/2020/05/16/the-real-threat-of-fake-voices-in-a-time-of-crisis/?guccounter=1`, 2020. Techcrunch; accessed 08-November-2020.

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

[566] F. Zhang, S. Zhou, Z. Qin, and J. Liu. Honeypot: a supplemented active defense system for network security. In *Proceedings of the Fourth International Conference on Parallel and Distributed Computing, Applications and Technologies*, pages 231–235. IEEE, 2003.

[567] L. Zhang and V. L. Thing. Three decades of deception techniques in active cyber defense-retrospect and outlook. *Computers & Security*, 106:102288, 2021.

[568] P. Zhang, W. Hui, B. Wang, D. Zhao, D. Song, C. Lioma, and J. G. Simonsen. Complex-valued Neural Network-based Quantum Language Models. *ACM Transactions on Information Systems (TOIS)*, 40(4):1–31, 2022.

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

[570] X. Zhang, D. Wu, L. Ding, H. Luo, C.-T. Lin, T.-P. Jung, and R. Chavarriaga. Tiny noise, big mistakes: adversarial perturbations induce errors in brain-computer interface spellers. *National Science Review*, 2020.

[571] 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.

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

[573] Z. Zhou, K. Xu, and J. Zhao. Homophily of music listening in online social networks of China. *Social Networks*, 55:160–169, 2018.

Curriculum Vitae

Dr. Nadisha-Marie Aliman, M.Sc.

Homepage: https://nadishamarie.jimdo.com/
Email: nadishamarie.aliman@gmail.com

Education

01/2021 – **Utrecht University**
Postdoctoral visiting scholar. Author and editor of books on Cyborgnetics.

03/2019 – 12/2020 **Utrecht University**
PhD degree in Computer Science.
(in approx. 1,5 instead of 4 years standard period of study)
▪ PhD Thesis: „Hybrid Cognitive-Affective Strategies for AI Safety"
▪ Specializations: AI Safety and Security, Cognitive Science, Affective Science and Affective Computing, Adversarial AI, Creativity, Meaningful Control of Intelligent Systems

04/2017 – 09/2018 **University of Stuttgart**
Master's degree in Computational Linguistics - summa cum laude
(in 3 instead of 4 semesters standard period of study)
▪ Master Thesis: „Cognitive Defense Mechanisms against Adversarial Examples in Natural Language Processing"
▪ Specializations: Applied Natural Language Processing, Cognitive Science, Adversarial Machine Learning, Machine Ethics, Emotion and Sentiment Analysis, Affective Computing

10/2014 – 03/2017 **Saarland University**
Bachelor's degree in Computational Linguistics
(in 5 instead of 6 semesters standard period of study)
▪ Bachelor Thesis:
„Ladder Networks for Named Entity Recognition"
▪ Supplementary subject: Computer Science
▪ Specializations: Deep Learning, Sentiment Analysis, Information Extraction, Security and Privacy

10/2009 – 06/2014 **University of Stuttgart and FH Kaiserslautern**
Orientation phase, modules out of diverse courses of study (i.a. „History of Science and Technology")

09/2006 – 06/2009 **Luisengymnasium Düsseldorf**
Abitur (in total shorter education through the skipping of 3 grades at elementary school)