Epistemic Defenses against Scientific and Empirical Adversarial AI Attacks*

Nadisha-Marie Aliman[1] and **Leon Kester**[2†]
[1]Utrecht University, Utrecht, The Netherlands
[2]TNO Netherlands, The Hague, The Netherlands
leon.kester@tno.nl

Abstract

In this paper, we introduce *"scientific and empirical adversarial AI attacks"* (SEA AI attacks) as umbrella term for not yet prevalent but technically feasible deliberate *malicious* acts of specifically crafting AI-generated samples to achieve an *epistemic distortion* in (applied) science or engineering contexts. In view of possible socio-psycho-technological impacts, it seems responsible to ponder countermeasures *from the onset on* and not in hindsight. In this vein, we consider two illustrative use cases: the example of AI-produced data to mislead *security engineering* practices and the conceivable prospect of AI-generated contents to manipulate *scientific writing* processes. Firstly, we contextualize the epistemic challenges that such future SEA AI attacks could pose to society in the light of broader i.a. *AI safety*, AI ethics and cybersecurity-relevant efforts. Secondly, we set forth a corresponding supportive generic *epistemic defense* approach. Thirdly, we effect a threat modelling for the two use cases and propose tailor-made defenses based on the foregoing generic deliberations. Strikingly, our transdisciplinary analysis suggests that employing distinct *explanation-anchored*, *trust-disentangled* and *adversarial* strategies is one possible principled *complementary* epistemic defense against SEA AI attacks – albeit with caveats yielding incentives for future work.

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 [Amodei *et al.*, 2016; Bostrom, 2017; Burden and Hernández-Orallo, 2020; Fickinger *et al.*, 2020; Leike *et al.*, 2017], AI ethics and AI governance issues [Floridi *et al.*, 2018; Jobin *et al.*, 2019; ÓhÉigeartaigh *et al.*, 2020; Raji *et al.*, 2020] gained momentum at an international level. In addition, cybersecurity-oriented frameworks in AI safety [Aliman *et al.*, 2021; Brundage *et al.*, 2018; Pistono and Yampolskiy, 2016] 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 [Zeadally *et al.*, 2020], AI simultaneously provides new affordances for attackers [Ashkenazy and Zini, 2019] from AI-aided social engineering [Seymour and Tully, 2016] to AI-concealed malware [Kirat *et al.*, 2018]. Next to the capacity of AI to extend classical cyberattacks in scope, speed and scale [Kaloudi and Li, 2020], 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 [Aliman *et al.*, 2021; Chesney and Citron, 2019; Kaloudi and Li, 2020; Tully and Foster, 2020] which is especially relevant to information warfare [Hartmann and Giles, 2020]. Recently, the weaponization of Generative AI for information operations has been described as *"a sincere threat to democracies"* [Hartmann and Steup, 2020]. In this paper, 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" [Zhao *et al.*, 2021] related to vision). However, for

†Contact Author

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 [Mahlangu *et al.*, 2019] 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) [Ranade *et al.*, 2021]. The authors utilized a fine-tuned version of the GPT-2 language model [Radford *et al.*, 2019] 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 [Wahle *et al.*, 2021] and with the pre-trained GPT-3 language model [Brown *et al.*, 2020] that *"may very well pass peer review"* [Dehouche, 2021] but also AI-generated fake reviews (with a fine-tuned version of GPT-2) apt to mislead experienced researchers in a small user study [Tallón-Ballesteros, 2020]. Future malicious actors could deliberately breed a large-scale agenda in the spirit of *"fake science news"* [Ho *et al.*, 2020] and AI-generated papers that would widely exceed in quality (later withdrawn) computer-generated research papers [Van Noorden, 2014] 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 3.1, the security engineering use case could e.g. involve dynamic domino-effects leading to large financial losses and even risks to human lives while the scientific writing use case seems to moreover reveal a *domain-general epistemic problem*. The *mere existence* of the latter also affects the former and could engender serious pitfalls whose generically formulated principled management is compactly treated in the next Section 2.

2 Theoretical Generic Epistemic Defenses

As reflected in the law of requisite variety (LRV) known from cybernetics, *"only variety can destroy variety"* [Ashby, 1961]. 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 [Vinnakota, 2013] to AI safety [Aliman, 2020a] over AI ethics [Ashby, 2020]. 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 [Wahle *et al.*, 2021] has been often thematized and even lately applied to "fake news" in the healthcare domain [Baris and Boukhers, 2021]. 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"* [Boneh *et al.*, 2019]. Another analysis presented a set of defense measures against the spread of deepfakes [Chesney and Citron, 2019] 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"* [Ho *et al.*, 2020], 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"* [Makri, 2017]. Following the author, scientists could equip citizens with sense-making tools without which *"emotions and beliefs that pander to false certainties become more credible"* [Makri, 2017].

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"* [Brown *et al.*, 2020]) 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"* [Floridi, 2018]. 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* [Fallis, 2020]. Thereby, it is assumed that *"deepfakes reduce the amount of information that videos carry to viewers"* [Fallis, 2020] 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 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 [Bufacchi, 2021] – 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 [Aliman and Kester, 2020]. 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"* [Popper, 1996]. Thirdly, the epistemic aim in science can neither be truth directly [Frederick, 2020] 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 [Popper, 2014]. Science is quintessentially *explanatory* i.e. it is based on explanations [Deutsch, 2011] and *not* merely on data. While the epistemic aim cannot be certainty or justification (and *not* even "truer explanations" [Frederick, 2020][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 [Frederick, 2020]. One can collectively agree on practical *updatable* criteria which better explanations should fulfill. In short, one does not assess a scientific theory in isolation, but in comparison to rival theories and one is thereby embedded in a context with other scientists. Fourthly, there are distinct ways to handle falsification and integrate empirical findings in explanation-anchored science. One can e.g. criticize an explanation 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 [Deutsch, 2016]. 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"* [Deutsch, 2016].

Against the background of this epistemic bedrock, one can now re-assess the threat landscape of SEA AI attacks. Firstly, one can conclude that AI-generated "fake data" and "fake experiments" could *slow down* but *not* terminally disrupt scientific and empirical procedures. In the case of misguiding confirmatory data, it has *no* epistemic effect since as opposed to empiricist epistemology, explanation-anchored science does not utilize any scheme of credence updates for a theory and it is clear that *"a severely tested but unfalsified theory may be false"* [Frederick, 2020]. 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"* [Frederick, 2020] if the accepted observations believed to have falsified it were wrong. Thirdly, when now considering the final case of AI-generated "fake reviews", it becomes clear that they could similarly *slow down* but *not* terminally disrupt the scientific method. At worst some existing theories could be unnecessarily problematized and misguiding theories uptaken, but all these epistemic procedures can be repealed retrospectively.

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"* [Ho *et al.*, 2020], 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* [Barrett, 2017]. 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

[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 [Frederick, 2020].

[2] That there could exist a task which imitative language models are *"theoretically incapable of handling"* has been often put into question [Sahlgren and Carlsson, 2021]. However, on epistemic grounds elaborated in-depth previously [Aliman, 2020a; Aliman *et al.*, 2021] which might be amenable to experimental falsifiability [Aliman, 2020b], we assume that the task to consciously *create and understand* novel yet unknown *explanatory* knowledge [Deutsch, 2011] – which humans are capable of performing *if willing to* – cannot be learned by AI systems *by mere imitation*.

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***

3 Practical Use of Theoretical Defenses

In the following Subsection 3.1, we briefly perform an exemplary threat modelling for the two specific use cases introduced in Section 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 [Aliman *et al.*, 2021]. In Subsection 3.2, we then derive corresponding tailor-made defenses from the generic characteristics that have been carved out in the last Section 2 while thematizing notable caveats.

3.1 Threat Modelling for Use Cases

Use Case Security Engineering

- ***Adversarial goals:*** As briefly mentioned in Section 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 [Ranade *et al.*, 2021]. 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"* [Ranade *et al.*, 2021]. 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 [Benz and Chatterjee, 2020; Ozkan *et al.*, 2021] 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 finetunes 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"* [Aliman *et al.*, 2021]. 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 [Mahlangu *et al.*, 2019] 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 [Prier, 2017; Hartmann and Giles, 2020]. 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 [Jakubowski, 2019]. Naturally, when logically applying the epistemic stance introduced in the last Section 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* [Carlini *et al.*, 2019], 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* [Aliman *et al.*, 2021] 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.

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 [Hartmann and Giles, 2020] and cyberattacks targeting AIs [Hartmann and Steup, 2020], consequences could have ranged from severe financial losses to threats to human lives. Multiple socio-psycho-technological solutions including the ones reviewed in Section 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 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* [Ranade *et al.*, 2021]. 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* [Heyvaert *et al.*, 2019] as signals to calculate when it will epistemically seek clarification from a 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*

[3] Such a system could instantiate *technical* self-awareness [Aliman, 2020a] (e.g. via active inference [Smith *et al.*, 2021]).

[4] The zero-trust [Kindervag, 2010] *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"* [Collier and Sarkis, 2021] seems highly appropriate in this increasingly complex security landscape.

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.

2. ***Trust-disentangled instead of trust-dependent:*** Via trust-disentanglement, a paper generated by a present-day AI would not only be rejected on provenance grounds but due to its merely imitative and non-explanatory content. Though, an important asset is the review process which if infiltrated by imitative AI-generated content could slow down explanation-anchored criticism if not thwarted fastly. A zero-trust scheme could mitigate this risk time-efficiently (e.g. via a consortium blockchain for review activities). Another zero-trust method would be to taxonomically monitor SEA AI attack events at an international level e.g. via an AI incident base [McGregor, 2020] tailored to these attacks and complemented by *adversarial* retrospective counterfactual risk analyses [Aliman *et al.*, 2021] 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 [Aliman *et al.*, 2021]. 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* [Kranenbarg *et al.*, 2018], one could perform red teaming, penetration tests and (adaptive) attacks employing AI-generated "fake data and experiments", "fake papers" and "fake reviews" [Tallón-Ballesteros, 2020]. Candidates for a blue team are e.g. reviewers and editors. Concurrently, urgent AI-related plagiarism issues arise [Dehouche, 2021].

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.

References

[Aliman and Kester, 2020] Nadisha-Marie Aliman and Leon Kester. Facing Immersive "Post-Truth" in AIVR? *Philosophies*, 5(4):45, 2020.

[Aliman *et al.*, 2021] Nadisha-Marie Aliman, Leon Kester, and Roman Yampolskiy. Transdisciplinary AI Observatory—Retrospective Analyses and Future-Oriented Contradistinctions. *Philosophies*, 6(1):6, 2021.

[Aliman, 2020a] Nadisha-Marie Aliman. *Hybrid Cognitive-Affective Strategies for AI Safety*. PhD thesis, Utrecht University, 2020.

[Aliman, 2020b] Nadisha-Marie Aliman. Self-Shielding Worlds. https://nadishamarie.jimdo.com/clipboard/, 2020. Online; accessed 23-November-2020.

[Amodei *et al.*, 2016] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

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

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

[Ashkenazy and Zini, 2019] Adi Ashkenazy and Shahar 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.

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

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

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

[Boneh *et al.*, 2019] Dan Boneh, Andrew J Grotto, Patrick McDaniel, and Nicolas Papernot. How relevant is the Turing test in the age of sophisbots? *IEEE Security & Privacy*, 17(6):64–71, 2019.

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

[Brown *et al.*, 2020] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[Brundage *et al.*, 2018] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.

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

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

[Carlini *et al.*, 2019] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

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

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

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

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

[Deutsch, 2016] David 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.

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

[Fickinger *et al.*, 2020] Arnaud Fickinger, Simon Zhuang, Dylan Hadfield-Menell, and Stuart Russell. Multi-principal assistance games. *arXiv preprint arXiv:2007.09540*, 2020.

[Floridi *et al.*, 2018] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca 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.

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

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

[Hartmann and Giles, 2020] Kim Hartmann and Keir 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.

[Hartmann and Steup, 2020] Kim Hartmann and Christoph 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.

[Heyvaert *et al.*, 2019] Pieter Heyvaert, Ben De Meester, Anastasia Dimou, and Ruben Verborgh. Rule-driven inconsistency resolution for knowledge graph generation rules. *Semantic Web*, 10(6):1071–1086, 2019.

[Ho *et al.*, 2020] Shirley S Ho, Tong Jee Goh, and Yan Wah Leung. Let's nab fake science news: Predicting scientists' support for interventions using the influence of presumed media influence model. *Journalism*, page 1464884920937488, 2020.

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

[Jobin *et al.*, 2019] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019.

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

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

[Kirat *et al.*, 2018] Dhilung Kirat, Jiyong Jang, and Marc Stoecklin. Deeplocker–concealing targeted attacks with AI locksmithing. *Blackhat USA*, 2018.

[Kranenbarg *et al.*, 2018] Marleen Weulen Kranenbarg, Thomas J Holt, and Jeroen 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.

[Leike *et al.*, 2017] Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. AI safety gridworlds. *arXiv preprint arXiv:1711.09883*, 2017.

[Mahlangu *et al.*, 2019] Thabo Mahlangu, Sinethemba January, Thulani Mashiane, Moses Dlamini, Sipho Ngobeni, Nkqubela Ruxwana, and Sun Tzu. Data Poisoning: Achilles Heel of Cyber Threat Intelligence Systems. In *Proceedings of the ICCWS 2019 14th International Conference on Cyber Warfare and Security: ICCWS*, 2019.

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

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

[ÓhÉigeartaigh *et al.*, 2020] Seán S ÓhÉigeartaigh, Jess Whittlestone, Yang Liu, Yi Zeng, and Zhe Liu. Overcoming barriers to cross-cultural cooperation in AI ethics and governance. *Philosophy & Technology*, 33(4):571–593, 2020.

[Ozkan *et al.*, 2021] Bilge Yigit Ozkan, Sonny van Lingen, and Marco Spruit. The Cybersecurity Focus Area Maturity (CYSFAM) Model. *Journal of Cybersecurity and Privacy*, 1(1):119–139, 2021.

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

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

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

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

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[Raji *et al.*, 2020] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily 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.

[Ranade *et al.*, 2021] Priyanka Ranade, Aritran Piplai, Sudip Mittal, Anupam Joshi, and Tim Finin. Generating Fake Cyber Threat Intelligence Using Transformer-Based Models. *arXiv preprint arXiv:2102.04351*, 2021.

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

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

[Smith *et al.*, 2021] Ryan Smith, Karl Friston, and Christopher Whyte. A Step-by-Step Tutorial on Active Inference and its Application to Empirical Data. *PsyArXiv*, 2021.

[Tallón-Ballesteros, 2020] AJ 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.

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

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

[Vinnakota, 2013] Tirumala 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.

[Wahle *et al.*, 2021] Jan Philip Wahle, Terry Ruas, Norman Meuschke, and Bela Gipp. Are neural language models good plagiarists? A benchmark for neural paraphrase detection. *arXiv preprint arXiv:2103.12450*, 2021.

[Zeadally *et al.*, 2020] Sherali Zeadally, Erwin Adi, Zubair Baig, and Imran A Khan. Harnessing artificial intelligence capabilities to improve cybersecurity. *IEEE Access*, 8:23817–23837, 2020.

[Zhao *et al.*, 2021] Bo Zhao, Shaozeng Zhang, Chunxue Xu, Yifan Sun, and Chengbin Deng. Deep fake geography? When geospatial data encounter Artificial Intelligence. *Cartography and Geographic Information Science*, pages 1–15, 2021.