10. Immoral programming

What can be done if malicious actors use language AI to launch 'deepfake science attacks'?

Nadisha-Marie Aliman[1] and Leon Kester[2*]

[1]Utrecht University, Faculty of Science, Department of Information and Computing Sciences, Princetonplein 5, 3584 CC Utrecht, the Netherlands; [2]TNO Netherlands, Intelligent Autonomous Systems, Postbus 96864, 2509 JG The Hague, the Netherlands; leon.kester@tno.nl

Abstract

The problem-solving and imitation capabilities of AI are increasing. In parallel, research addressing ethical AI design has gained momentum internationally. However, from a cybersecurity-oriented perspective in AI safety, it is vital to also analyse and counteract the risks posed by intentional malice. Malicious actors could for instance exploit the attack surface of already deployed AI, poison AI training data, sabotage AI systems at the pre-deployment stage or deliberately design hazardous AI. At a time when topics such as fake news, disinformation, deepfakes and, recently, fake science are affecting online debates in the population at large but also specifically in scientific circles, we thematise the following elephant in the room now and not in hindsight: what can be done if malicious actors use AI for not yet prevalent but technically feasible 'deepfake science attacks', i.e. on (applied) science itself? Deepfakes are not restricted to audio and visual phenomena, and deepfake text whose impact could be potentiated with regard to speed, scope, and scale may represent an underestimated avenue for malicious actors. Not only has the imitation capacity of AI improved dramatically, e.g. with the advent of advanced language AI such as GPT-3 (Brown *et al.*, 2020), but generally, present-day AI can already be abused for goals such as (cyber)crime (Kaloudi and Li, 2020) and information warfare (Hartmann and Giles, 2020). Deepfake science attacks on (applied) science and engineering – which belong to the class of what we technically denote as scientific and empirical adversarial (SEA) AI attacks (Aliman and Kester, 2021) – could be instrumental in achieving such aims due to socio-psycho-technological intricacies against which science might not be immune. But if not immunity, could one achieve resilience? This chapter familiarises the reader with a complementary solution to this complex issue:

Bart Wernaart (ed.)
Moral design and technology
DOI: 10.3920/978-90-8686-922-0_10, © Wageningen Academic Publishers 2022

a generic 'cyborgnetic' defence (GCD) against SEA AI attacks. As briefly introduced in Chapter 4, the term cyborgnet (which is much more general than and not to be confused with the term 'cyborg') stands for a generic, substrate-independent and hybrid functional unit which is instantiated e.g. in couplings of present-day AIs and humans. Amongst many others, GCD uses epistemology, cybersecurity, cybernetics, and creativity research to tailor 10 generic strategies to the concrete exemplary use case of a large language model such as GPT-3. GCD can act as a cognitively diverse transdisciplinary scaffold to defend against SEA AI attacks – albeit with specific caveats.

Key concepts

- For safety reasons, it is vital to tackle the immoral programming issue of intentional malice
- Malicious actors could launch deepfake science attacks against the science enterprise
- Science is not immune to such attacks, and proactive defences are required
- Generic 'Cyborgnetic' Defence (GCD) is a transdisciplinary framework that crafts solutions from a cyborgnetic stance
- A cyborgnet is not to be confused with a cyborg. A cyborgnet is a generic substrate-independent hybrid functional unit (i.e. all cyborgs exist in cyborgnets but not the reverse)
- GCD is a complementary defence against deepfake science (or technically SEA AI) attacks
- Thereby, GCD acts as a cognitively diverse scaffold that uses epistemology, creativity research and knowledge from many other fields. GCD could be resilient – but not immune

10.1 The practical scientific and empirical adversarial AI attack problem

The not yet prevalent but technically feasible scientific and empirical adversarial (SEA) AI attacks could be launched in multiple modalities. However, for illustrative purposes, we focus on text-based SEA AI attacks using language models. We analyse three attack vectors: (1) AI-generated data and experiments, (2) AI-generated research articles, (3) AI-generated reviews. Firstly, it is noteworthy that the idea to artificially generate academic text contributions and inject them even into respected venues has been already implemented in some (later withdrawn) cases (Van Noorden, 2014) merely on the basis of a mediocre automated text generation mechanism. As researchers discovered (Eckert *et al.*, 2018), this was similarly possible with human-generated made-up empirical studies accepted by predatory publishers to which internationally

AI-generated data and experiments

respected scientists were found to nevertheless habitually submit contributions. With language models, malicious actors could potentiate such practices with regard to superficial linguistic quality, speed, scale, and scope including pre-print proliferations. In addition, in applied science or engineering contexts such as in cybersecurity, the emergence of sophisticated language models such as GPT-2 (Radford *et al.*, 2019) and GPT-3 (Brown *et al.*, 2020) open up an unprecedented field of affordances for malicious actors. For instance, recently, it has been shown that AI-generated cyber threat intelligence (CTI) obtained via a fine-tuned version of GPT-2 could be utilised to poison data-driven cyber defence systems at training time (Ranade *et al.*, 2021). This could have serious security consequences, given the growing cybercrime damages expected to reach 6 trillion USD in 2021 (Ozkan *et al.*, 2021) and the associated risk for critical infrastructure. Importantly, the abovementioned artificially generated text CTI samples, which reported about distorted cyber threat events, were also able to fool human cybersecurity experts who 'labelled the majority of the fake CTI samples as true despite their expertise' (Ranade *et al.*, 2021). Generally, legitimate experimental research relying on online data could be sabotaged on a large scale via such deepfake-based poisoning schemes. (A related feasible and serious but not yet prevalent recent concern from a very different science field is deepfake geography (Zhao *et al.*, 2021), i.e. AI-generated fictional satellite images in GIScience.) Beyond that, in a tentative prompting of a publicly available interface to the pre-trained GPT-2 model (Radford *et al.*, 2019), we found that the model is able to generate text samples vaguely mimicking academic paper structures and – for example – referring to never conducted experiments (Figure 10.1). While the consistency was often low, note that the utilised GPT-2 model has been neither fine-tuned on research papers nor is it as powerful as its successor GPT-3, whose parameters are two orders of magnitude bigger (Bartoli and Medvet, 2020) and which still has a closed access policy.

AI-generated research articles

Secondly, regarding more theoretical AI-generated research articles, the early GPT-3 paper already touched upon that potential misuse, albeit only briefly, via the mention of the expression 'fraudulent academic essay writing' (Brown *et al.*, 2020). In a very recent plagiarism study with the pre-trained GPT-3, it has been stated that in view of the quality of the textual results it 'may very well pass peer review' (Dehouche, 2021). Malicious actors could utilise such AI-generated contributions as instruments for information operations. For instance, in order to reinforce the sentiment of unresolvable disagreement also in the science field, they could fuel narratives on 'post-truth' (which, as mentioned in Chapter 4, we consider to be a misguiding misnomer) via a novel form of scientific astroturfing (Aliman and Kester, 2021). Thereby, astroturfing (Hartmann and Giles, 2020) is a deceptive online technique deliberately utilised in order to provide the impression that a certain asset is widely supported or

Figure 10.1. Counterfactual outputs of pre-trained GPT-2.

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

rejected. This in turn combined with the societal-level automated disconcertion phenomenon (Aliman *et al.*, 2021). For instance, the fundamental epistemic confusion that arose by the mere existence of deepfakes may be able to trigger destabilising processes in fragile societies – one of the main aims in information warfare. Conversely, malicious actors could also seek to automate a flood of corroborative papers confirming the efficiency of an application, justifying a certain theory or the robustness of a defence method to exploit the vulnerabilities that a potentially decreased security awareness could bring about. The latter

could in turn lead to domino effects related to the cybersecurity risks mentioned in the last paragraph. Apart from that, in our tentative probing of the publicly accessible pre-trained GPT-2, the model was able to output a few text blocks of passable quality on specific topics which could be assembled for abstracts (for a few examples, see Figure 10.1). Moreover, among others, the following artefacts that were generated may be of interest for future work: fictional links and references to fictional quotes attributed to individuals with synthetic names, self-generated structures for sections and even acknowledgment sections with mention of existing or fictional research institutes and specification of synthetic grants. Finally, we fed the twofold title of this very chapter into an interface for GPT-Neo (1.3B) (Eleuther AI, 2021), an open-source GPT-inspired language model trained on a dataset denoted 'The Pile' (Gao *et al.*, 2020). This dataset contains, among others, a large number of scientific papers and abstracts. The interface restricted its output to a certain number of characters (around the length of a sentence). The model outputted the following string: 'Intrusion Detection Systems (IDS) can detect potentially dangerous behaviours such as malicious actors launching deepfake science attacks disguised as science'. We leave this AI-generated text sample without comment for now and will return to it in Section 10.3. However, it is already worth stating that, generally, an explanation-anchored criticism is content-centred and does not need to consider the source (see Section 10.2). After the submission of this very book chapter, Eleuther AI (2021) released the novel open-source language model GPT-J (6B) which is larger than GPT-Neo (1.3B) (but not GPT-3) and has also been trained on the mentioned Pile dataset (Gao *et al.*, 2020). Strikingly, the bilingual language model Wu Dao 2.0, which is even ten times larger than GPT-3, has (among others) also been trained on the Pile (Romero, 2021). In light of the results from our tentative probing of the publicly accessible GPT-J (6B) model (displayed in Figure 10.2), we conjecture that a comparable or superior large language model trained uniquely (and not only partially) on scientific and technical papers may significantly boost the capabilities of SEA AI attackers. However, as will become apparent in Subsection 10.2.2., the same type of tool could conversely facilitate powerful defence strategies against such attacks and beyond that augment scientific creativity – a potentially safety-relevant quest for humanity. Hence, future work could investigate further along those lines.

AI-generated reviews

Thirdly, coming to the topic of AI-generated reviews, a small user study (Bartoli and Medvet, 2020) found that it was possible to mislead experienced human reviewers with text samples obtained from a fine-tuned version of GPT-2. Given the length of reviews, it seems that such settings could appear lucrative for malicious goals including insider threats, for example in the presence of financial motives. Crucially, reviewers have been described as gatekeepers of knowledge (Tsao *et al.*, 2019) that can influence the progress of creative ideas

Figure 10.2. Counterfactual outputs of pre-trained GPT-J 6B.

Prompt by human	Counterfactual outputs of pre-trained GPT-J 6B
Abstract In a time in which topics such as fake news, disinformation, deepfakes and recently fake science affect online debates amounts the population at large, but also specifically in scientific circles, we thematize the following elephant in the room now and not in hindsight: what to do if malicious actors use AI for not yet prevalent but technically feasible 'deepfake science attacks' i.e., on (applied) science itself? Deepfakes are not restricted to audio and visual phenomena and deepfake text whose impact could be potentiated with regard to speed, scope, and scale may represent an underestimated avenue for malicious actors.	[...] The occurrence of such attacks not only influences societal discourse and thus the public opinion of a society, but also impacts the public acceptance of the findings of academic science (Keil, 2019). In particular, deepfake science attacks are very easy to create (Kim, 2016). [...] We wish to convey our urgent call for scientists to be attentive of possible malicious AI/AS scenarios [...] The rapid increase of deepfake science could considerably harm a number of fields and even drive a number of current and future sciences to become incomprehensible to society. It also raises serious ethical and legal questions concerning scientific integrity, data privacy, and/or copyright. [...] we state that the advent of unsupervised machine learning for 'deepfake science' poses a growing threat [...] [...] The attacks presented in this paper show how AI is now used in text manipulation to alter and attack human perceptions of a scientific document. They do not make use of fake videos, fake images, fake audio, fake faces, fake objects, or fake messages, but manipulate text. Furthermore, we want to show that even though these attacks are in the scope of deepfake science and its sub-topic of deepfake text, their goal is to influence the public discourse. In this paper, we intend to exemplify such threat by investigating whether these fake news-like approaches will have an impact on scientific decision-making, trust, and decision credibility, as well as have an impact on the researcher's and academic institution's career or even lead to possible arrest and court trials. Our results are still early, yet they seem to be telling. [...] The overall paper focuses on all the technical aspects that come along with manipulating texts or data that include theoretical approaches, algorithms, models, tools and technology, with the ultimate goal of enabling an AI detection system. The contribution is in two directions: (1) how to mitigate and stop deepfake science, and (2) what kind of AI will manipulate science.

in a society. However, through automated disconcertion, a reviewer could in principle always object that a paper was presumably AI-generated. Also, if not counteracted early, an increased fraction of AI-generated reviews could skew the directions that science takes in the long term with many socio-psycho-technological repercussions. Furthermore, reviewers that act on empiricist epistemologies that seek for the justification of truer beliefs via probabilistic belief updates may face (unnecessarily in our view, as implied in Section 10.2)

'epistemic threats' (Fallis, 2020). On this score, the amount of information in audiovisual material decreases steadily due to widespread deployment of deepfake videos (Fallis, 2020), while a similar quantitative impact analogously already affected text material even earlier via fake news. An increased awareness in society with regard to such epistemically relevant problems and confusions in science could in turn exacerbate automated disconcertion and be instrumental in information operations as described earlier. To sum up, while some scientists may at first sight be under the impression that they are immune to such purposeful text-based immoral programming with the goal of provoking an AI-aided epistemic distortion (either as an end in itself or instrumental in achieving further malicious aims), a deeper analysis suggests that unfortunately this may not always be the case.

Exemplary text segments from our probing of the publicly available GPT-2 application interface accessible at: https://deepai.org/machine-learning-model/text-generator. The fragments were sampled from 20 consecutive prompts and have been hand-chosen to illustrate some of the extracted features as discussed in the text. The outputs are not deterministic and can vary widely in linguistic quality and consistency. However, 'some meta-cherry picking' (Radford *et al.*, 2019) has been as well performed by Open AI itself when displaying abilities of GPT-2 for demonstration purposes. The human prompt indicated corresponds to the conjunction of 'Abstract', a newline character and two sentences sampled from the first page of this very chapter on immoral programming.

Exemplary text segments from our probing of the publicly available GPT-J 6B application interface accessible at: https://6b.eleuther.ai/. The fragments were sampled from 20 consecutive prompts and have been hand-chosen to illustrate some of the extracted features as discussed in the text. The outputs are not deterministic and can vary in linguistic quality and consistency depending on the chosen parameters.

10.2 Generic cyborgnetic defence as complementary theoretical solution

Against the backdrop of the above-described possible SEA AI attack scenarios, this section introduces Generic 'Cyborgnetic' Defence (GCD) framed as a countermeasure to such attacks. GCD provides a novel (unquestionably non-exhaustive and hence to be steadily updated) set of generic strategies formulated from a cyborgnetic stance. The concept of a cyborgnet was previously introduced in Chapter 4. On an inflationary account extending beyond the mere study of systems, so-called Type II entities are all those for which it is

possible to consciously create and understand explanatory knowledge, while Type I entities are all entities for which this is conjectured to be impossible[5]. Based on that, a cyborgnet represents a generic, substrate-independent and hybrid functional unit comprising relations between Type II entities (so today only applicable to humans including, but not restricted to, cyborgs) and Type I entities (such as Type I AIs but also any other Type I entities not limited to systems and thus also, e.g. ideas, processes, or objects). On that view, early Type II humans equipped with complex material tools and language abilities already had an inherently cyborgnetic existence with both material and linguistic tools representing integrated Type I entities (Aliman, 2021a). Today, not long after the intricacies of deepfake and automated disconcertion started to affect the information ecosystem, SEA AI attacks could now become the entry point for analogously discombobulating phenomena in the scientific ecosystem. While this can seem threatening to scientists, it need not be – if science creatively adapts to this novel complex field of affordances whilst not interrupting its quest for better explanations. In this vein, one motivation for both an epistemic and a creativity-centred cyborgnetic stance for a generic SEA AI attack defence, is the law of requisite variety from cybernetics stating that 'only variety can destroy variety' (Ashby, 1961). Since the malicious adversary operates from within a coupling with a language model targeting the victim at an epistemic level, a defender may profit from integrating not only epistemic knowledge but also such knowledge stemming from language models too. Building on this, the generic defences under GCD aim at: (1) facilitating resilience to malicious actors and their language models while (2) simultaneously facilitating a creativity-augmenting feedback loop between defenders and their own language models. The former is an inter-cyborgnetic and the latter an intra-cyborgnetic endeavour. While perhaps unintuitive since potentially unusual at first sight, we offer a deeper explanation of this line of thought, to which we return in the next subsection.

10.2.1 Generic epistemic defence

From a functional cyborgnetic point of view, an extremely vital asset in security contexts is embodied cyborgnetic creativity. Since one is not able to reduce the adversarial disturbances in the form of SEA AI attacks controlled

[5] Note that this ontology (Aliman, 2020; Aliman, 2021a) has no relation whatsoever to the metaphor of Kahneman related to 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 (Hutchinson and Barrett, 2019)). Conversely, currently known Type II entities are restricted to humans as a species and examples for Type I entities are everything else. This means Type I entities can be e.g. non-human conscious mammals like dogs, but also thoughts, language itself, mechanical tools, dreams, decision trees, chatbots, etc.

by malicious actors being unpredictable explanatory knowledge creators, a risk averse solution cannot be the only option. In fact, instead of shielding oneself from deepfake texts, which could in the long term even necessitate a retreat from society, another strategy could consist of 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 (Fallis, 2020) 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 (Aliman and Kester, 2021), 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 (Fallis, 2020)). The aim is not to find isolated good explanations, but to identify better ones (Frederick, 2020) according to criteria agreed upon with others.

Hence, the first generic epistemic defence against SEA AI attacks is to select an explanatory-anchored approach to science instead of the prevailing data-driven one. A crucial advantage of explanatory-anchored science is its concurrently open-minded nature with regard to the momentary primary uptake of ideas in order to be able to inspect and criticise them but also its self-shielding nature when it comes to the second step of a permission for that idea to provisionally stay in one's prior web of knowledge being filtered by explanatory knowledge – which in turn can precisely not be mimicked by Type I AI. Explanation-anchored science is thus also its own defence method in the face of SEA AI attacks. The second generic defence against SEA AI attacks within GCD is a trust-disentangled approach that divorces content from source. In this way, much less importance would be assigned to deepfake detection endeavours embedded in incessant cat and mouse games. Ideally, the integrity of explanation-anchored messages can be afforded by the content of the messages themselves as if connected via an invisible blockchain. In other social settings disjunct from science, people may have multiple reasons why a

shielding from Type I entities stemming from their own or from other cyborgnets is desirable. For instance, a deceptive one-sided romantic relationship between a person and a chatbot can seem unwelcome. Similarly, one might try to avoid any involuntary investment of time on social media debates with future sophisticated bots acting as trolls or decide to forestall automated social engineering attempts in future social virtual reality. For all these practical cases, it is reasonable to implement something akin to a substrate-independent Type I shield (which is not to be confused and would not be equivalent to a Turing Test, see e.g. (Aliman, 2021a) for more details) if possible. However, in the science domain, there is as such no fundamental reason to shield oneself from ideas that one interprets from outputs generated by Type I AI.

There is no logical reason to assume a priori that everything generated by a Type I AI must necessarily be false. Even human liars are fallible and thus able to mistakenly tell something that may be true or stimulating in creative ways. So could the output of a Type I AI by chance sometimes contain some elements that humans might interpret as thought-provoking. Trust-disentanglement accommodates for that by allowing for novelties in deliberate and spontaneous idea generation. At worst, the Type I-generated content is rejected since both non-explanatory and useless. At best, the non-explanatory output comes with an additional element that stimulates creativity or a criticism of one's best present explanations – which brings us to the third and last generic epistemic defence under GCD: adversarial science. As already touched upon in Chapter 4, an adversarial approach to one's best prior conjectures is a rational creativity-stimulating strategy since one might e.g. unpredictably be able to falsify them and discover novel candidate better explanations upon acting against the old ones (Frederick, 2020). Pre-eminently, this signifies that explanatory-anchored science is not bound in any way to act on its best available explanations. This is decisively different from classical approaches such as encountered in empiricist and utility maximisation schemes that operate according to a fixed formula containing a set of options to which one is epistemically bound, which ignores spontaneous unpredictable creativity. In the main, explanatory-anchored adversarial science applies an adversarial paradigm to itself to such an extent that theories are purposefully formulated in a risky fashion such that they could be potentially and easily falsified. This allows for fast updates of knowledge and helps to avoid greater practical damages that could emerge by a prolonged stagnation in misleading assumptions. The goal is not to embellish one's conjectures and try to formulate them as carefully as possible to escape criticism. The aim is to formulate strong bold universal statements (Frederick, 2021). As stated by Popper, the more a theory forbids, the better that theory is (Popper, 1963). This leads us back to the beginning of this subsection. It now becomes clear why from an epistemic perspective, a self-paced exposure to

adversarial patterns combined with creativity-augmenting measures may be helpful in building resilience to SEA AI attacks involving inter-cyborgnetic and intra-cyborgnetic feedback loops. In Section 10.3, we address the question of how to implement such generic strategies in practice. Prior to that, the next subsection first introduces compatible generic methods for cyborgnetic creativity augmentation in a pragmatic framework, compiling insights from creativity research in the fields of psychology and cognitive neuroscience, i.e. now formulated from a scientific and empirical stance.

10.2.2 Generic cyborgnetic creativity augmentation

The ambiguously designated artificial creativity augmentation research direction (Aliman, 2020) has recently been put forth for the purpose of implementing generic defences 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. It seems well suited as a basis for crafting synergetic enhancement strategies for the intra-cyborgnetic feedback loop between human defenders and their language models. Applied to our generic defences against SEA AI attacks supported by language models, the twofold task can be exemplarily reformulated as follows: (1) augmenting human creativity using language models; and (2) augmenting artificial creativity in language models via humans. The former and the latter are intertwined since the subtask: (1) can reinforce the subtask; and (2) *vice versa*. In the spirit of recent work by Mick Ashby (2020) at the intersection of cybernetics and AI ethics, one could state that in this case, humans and language models reciprocally become a sort of 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. Compellingly, 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 language models. The second cluster pertains to generic strategies for the augmentation of artificial creativity within language models. To this end, we select suitable starting points based on the ten provisional available artificial creativity augmentation indicators (Aliman, 2020) which were grounded in explanations from psychology and cognitive neuroscience. On this score, seven

possible indicators suggested to enhance human creativity were: transformative criticism and contrariness, divergent thinking training, alteration of waking consciousness, active forgetting (during sleep), frequent engagement, brain stimulation as well as sensory extension. Moreover, three indicators suggested to enhance artificial creativity were: immersion in the human affective niche, social cognition, and an egocentric integrated multimodal virtual reality experience of the world. In this paper, we focus on those strategies that are technically implementable in present-day advanced language models. Thus, we limit our analysis to the first six indicators specified for human creativity enhancement and to the first indicator mentioned for the augmentation of artificial creativity. Firstly, in order to augment human creativity using language models, suitable generic strategies could be to design these AIs with the following enhancing subgoals: (1) increase human criticism abilities; (2) stimulate human divergent thinking; (3) alter the nature of self-experience at waking time; (4) extend the nocturnal unconscious and/or dream-related creative generation and active forgetting processes; (5) encourage frequent human engagement; (6) provide human sensory extension. Secondly, concerning the human-performed augmentation of artificial creativity within language models, we add the following generic strategy; (7) immersion in the human affective niche via a mathematical approach and via active sampling. These seven generic strategies against SEA AI attacks via cyborgnetic creativity augmentation can seem abstract at first sight. For this reason, the next Section 10.3 now instantiates and illustrates their application (together with the three generic epistemic defence strategies from the last subsection) using design fictions for the use case of large language models – the same tools that malicious actors could utilise for advanced SEA AI attacks.

generic strategies

10.3 Practical use of theoretical solution

In Section 10.2, we introduced the reader to our GCD framework consisting of three generic epistemic defence strategies and seven generic cyborgnetic creativity augmentation strategies against SEA AI attacks. For illustrative purposes, we now apply those methods to the practical large language model (abbreviated by LLM in the following) use case. We use design-fictions as recommended in AI safety frameworks (Aliman *et al.*, 2021). In this section, we see how one key trick in applying the GCD framework to practically relevant defences against SEA AI attacks performed with an LLM, is to generate desirable upward counterfactuals of a possible defence with GPT-3 itself. Step-by-step, we systematically proceed through all three practically relevant SEA AI attack vectors specified in Section 10.1: (1) AI-generated data and experiments; (2) AI-generated research articles; (3) AI-generated reviews. For each attack vector, we clarify how instances of generic epistemic defences and generic cyborgnetic creativity augmentation measures can help in practice.

Since comparable practical reports on LLMs such as GPT-3 are still relatively scarce, we rely on plausible design-fictions of what we already deem technically feasible today, i.e. we craft upward counterfactuals projecting to the immediate past (e.g. literally yesterday). Normally, it makes sense to project such design-fictions to the immediate future. However, we specifically frame it in this way to stress the attainability of many potentially valuable opportunities. Upward counterfactuals pertain to better ways in which scenarios could have unfolded but did not. This means we conjecture a world in which large language models do not undergo a closed source policy and any interested entity (such as a scientist) with reasonable resources could have acquainted itself with an LLM interface. For simplicity, given the complexity of the underlying issue, we assume that this entity is able to design novel applications for the LLM and to modify the model (e.g. to fine-tune it on other datasets – although this might not always be necessary anymore –, change its loss function and parameters or to extend it with other available technologies). Moreover, to simplify, we assume that the human SEA AI attacker appears to the defending scientist as a grey box instance with the only information being that the attacker owns an LLM too. Also, within attack-defence cycles, the conjunction of scientist and LLM instantiate a cyborgnet as does the conjunction of attacker and LLM. A simplified illustration of important intra-cyborgnet and inter-cyborgnet relations are depicted in Figure 10.3. Whilst both attacker and defender are naturally embedded in a complex heterogeneous and multi-layered socio-psycho-techno-physical environment and while the cyborgnets of each of them can contain a much larger number of Type I entities (e.g. ranging from ideas to technologies over processes) in a given situated conceptualisation, we abstract away further details for a better overview and for the purpose of a better visualisation. Also, from a cyborgnetic stance, even before language models like GPT-3, in fact, since the advent of human linguistic abilities, humans use language as a form of technological tool since it involves the application of explanatory knowledge for practical aims such as teaching, learning and participatory sense-making (Aliman, 2020; Aliman and Kester, 2021). In brief, language models add new nested dimensions to the linguistic tools in one's cyborgnet.

10.3.1 Cyborgnetic defence against hypothetical LLM-generated data and experiments

Concerning epistemic defences, one could have implemented the following. From the perspective of engineers in security, AI, but also in many other domains as well as scientists involved in empirical studies, a first step could have been to consciously familiarise oneself with the different steadily shifting cyborgnet constructs at different spatiotemporal scales in which one is embedded, while performing research with data and different systems. The

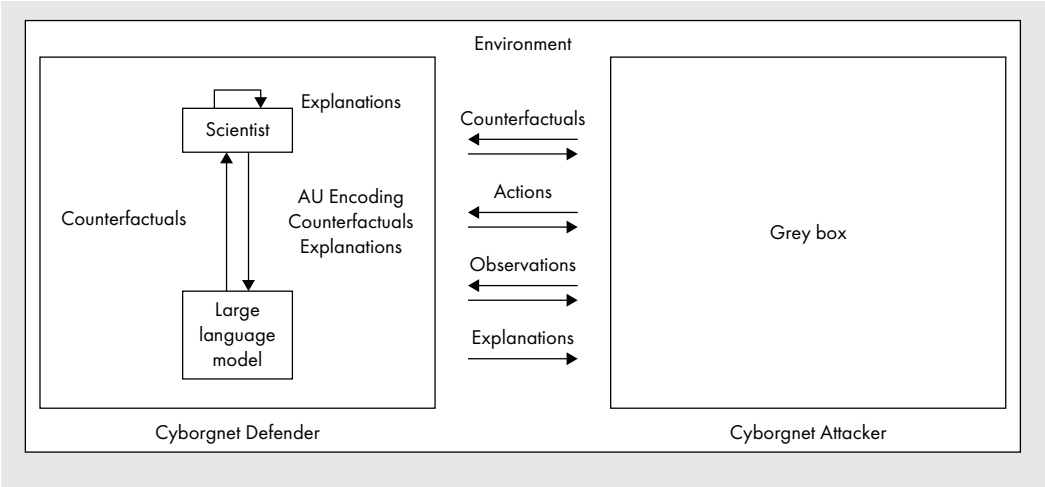

Figure 10.3. Simplified illustration and contextualisation of GCD-relevant intra- and inter-cyborgnet relations embedded in cycles of SEA AI attacks and defences performed by the cyborgnet of the attacker and the cyborgnet of the defender. Obviously, the entity labelled scientist could just as well be an engineer or a researcher from other areas.

cyborgnet inventory analysis

SEA AI attack surface is inherently cyborgnetic and could be much wider than it might seem with text inputs potentially hidden in non-salient parts of the research pipeline. It is feasible that certain SEA AI attacks may even stay undetected with only the repercussions becoming perceptible. Hence, one could have attempted to sketch individual- and group-level cyborgnet inventory maps in analogies to red team versus blue team settings (Rajendran *et al.*, 2011) to investigate how and via which assets and backdoors a covert textual SEA AI attack could potentially manifest itself. In such a cyborgnet inventory analysis, one could have specified underlying relations and feedback loops as minimally illustrated in Figure 10.3. Regarding the submission of one's own experimental contributions against the backdrop of LLM-generated samples, one could have adopted a refined strategy. Instead of mainly focusing on the experimental results, one could have strived for a thorough theoretical foundation coalescing theoretical and empirical elements. Every empirical study could have been supplemented with one explanatory section in which the empirical approach is contextualised against a larger scientific and theoretical background. Experiments designed to merely corroborate a theory could have been discarded. The scientific community could have opted for registered reports (West and Bergstrom, 2021), which represents a solution in which experimental research is assessed at an earlier stage based on the explanatory quality of the research proposal itself and not on the actual experimental results. Thus, one could have strived for exclusively explanation-anchored research proposals to which one could have in addition adopted an adversarial stance in order to improve them. One could have shifted the focus away from fixed

measures on pre-determined data and replaced a normative design philosophy striving for highly intelligent Type I systems instead to ever more creative embodied cyborgnets without any fixed performance measure. One's socio-psycho-techno-physical state could have been synergistically entangled with the system. (Logically speaking, it is not because ethical or aesthetic goals and problem-solving ability are separable that one needs to keep them separated. In Chapter 4, we explained why foregrounding embodied cyborgnetic creativity may mitigate the risks of advanced Type I AI control.) Finally, one could have accommodated epistemic dizziness in experimental procedures and while being safety-aware, one would also have accepted that the future of cyborgnet safety and security cannot be predicted.

With respect to cyborgnetic creativity augmentation, this paragraph discusses how LLMs could have been used to augment the creativity of humans engaging in empirical research. Creativity can be described as a tripartite evolutionary affective construct with three modes (Dietrich, 2019): 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. LLMs could have been utilised frequently to stimulate divergent thinking in the deliberate mode by first letting the scientist prompt the LLM on providing a solution to a given practical problem. Since LLMS are not able to create explanatory knowledge, the scientist could then criticise the generated output and re-prompt the LLM, derive inspiration from it, or utilise it to question own prior assumptions. By way of example, let us consider the output generated by GPT-Neo when prompted with the title and subtitle of this chapter. Namely, the answer was: 'Intrusion Detection Systems (IDS) can detect potentially dangerous behaviours such as malicious actors launching deepfake science attacks disguised as science'. We return to an LLM-aided critical assessment of the content of that statement in a few sentences. Generally, to improve the required critical reasoning abilities, a novel LLM-based systematic adversarial educational tool could have become available in empirical research. The LLM could have been utilised for life-long learning and for students in engineering and science to train the formulation of better explanation-anchored empirical research proposals, e.g. for the abovementioned registered reports. For instance, given a current paragraph and a history of earlier paragraphs, a student's next paragraph could then have competed with the LLM-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 LLM-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 invisible explanatory blockchain so to speak. The second aim could have been 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 have thereby been implemented due to the student being exposed to the alternative LLM-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' (Fink *et al.*, 2010). Interestingly, the 'other' in this case, while not being an explanatory knowledge creator, could have been the LLM, and the group-based functional unit could have been the cyborgnet. The LLM in turn could have been enhanced by fine-tuning the student's inputs at a later stage. Hence, this educational tool could have been called adversarial cyborgnetic cognitive stimulation. Coming back to the output of GPT-Neo on an intrusion detection system for deepfake science, one could then have investigated whether adversarial cyborgnetic cognitive stimulation (combined with a normalisation smoothing out superficial linguistic style differences) could allow for a subtly different defensive scheme with a similar effect: an explanatory intrusion prevention system (IPS) for science (Aliman, 2021b). Such a shielding IPS preceding scientific peer review could have been combined with a substrate-independent Type-I-shield or, technically, a substrate-independent Type-I-falsification-event test[6] (Aliman, 2021a). Its goal could have been to shield from non-explanatory texts – vitally however, without being equivalent to a deepfake science detection system. Thereby, such an explanatory IPS could not have been fully automated using a human evaluator.

adversarial cyborgnetic cognitive stimulation

explanatory intrusion prevention system

10.3.2 Cyborgnetic defence against hypothetical LLM-generated research papers

In connection with epistemic defences against SEA AI attacks utilising LLM-generated research papers, the just depicted adversarial cyborgnetic cognitive stimulation could have been proactively employed by scientists for self-education and life-long learning to improve explanation-anchored scientific writing practices. Furthermore, scientists could have engaged in red teaming and penetration testing procedures injecting LLM-generated papers into

[6] Such a substrate-independent Type-I-falsification-event test (Aliman, 2021a) requires a Type II evaluator (so specifically, a human nowadays) and merely leads to two asymmetric clusters: a first homogenous Type-I-free cluster and a second potentially heterogeneous cluster which, next to Type I entities, can also comprise Type II entities that have not yet passed the test (for instance because no suitable knowledge area tailored to the Type II test subject has been identified, because the Type II subject is still too young, for lack of motivation or willingness on the part of the Type II entity and so forth). In short, it is formally very different from the widespread idea of Turing Tests.

the submission process. To responsibly implement such schemes, scientists could have worked out tailored coordinated vulnerability disclosure practices (Kranenbarg *et al.*, 2018). For an LLM-aided trust-disentanglement to counteract SEA AI attacks at submission time, scientists could have experimented with the explanatory IPS tool just mentioned. Instead of striving for deepfake detection techniques, scientists could have aimed at implementing a scheme in which contents are not rejected because of the source that submitted them, but purely on explanation-anchored grounds. From this perspective, deepfake science papers would not have passed through the explanatory IPS because they have been generated by a Type I entity, but because those papers are merely of an imitative and hence non-explanatory nature. Pre-print platforms could have combined an automated active sampling of newly uploaded papers with an explanatory IPS involving human evaluators and LLMs. With regard to cyborgnetic creativity augmentation measures against SEA AI attacks, LLMs could have been 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 utilised as a 'multiversal' language model (Reynolds and McDonell, 2021), interactively generating branches of fictional counterfactuals to stimulate human creativity in fictional writing. Extending beyond that, scientists could now have combined an LLM-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 to experimentally falsify statements of interest that would have guided the extension of counterfactual nodes.

multiversal cyborgnetic co-creation

This multiversal cyborgnetic co-creation could have been further fine-tuned by scientists. Firstly, one could have had increased the immersion of the LLM 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. Interestingly, scientists could have used the input-agnostic generic mathematical scaffold and encoding of augmented utilitarianism (AU) (introduced in Chapter 4) to specifically tailor such parameters for the LLM they own. This conceptual idea is reflected in Figure 10.3 with the arrow labelled 'AU-encoding', flowing from scientist to LLM in the cyborgnet of the defender. Secondly, while language models like GPT-3 are imitative, outcomes perceived as creative are mainly those that exhibit implausible utility (Tsao *et al.*, 2019), i.e. utile outcomes with unexpectedly surprising previously underestimated facets. Scientists in their quest for implausible utility, could have been inspired by the idea of transdisciplinary cross-pollination effects and insights from research on cognitive diversity (Mitchell *et al.*, 2017; Reynolds and Lewis, 2017). 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 LLM, scientists could have been motivated by composer-audience architectures (Bunescu and Uduehi, 2019) from computational creativity (Franceschelli and Musolesi, 2021) utilised 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 pre-trained audience model outputs for that dataset. Analogously, scientists could have used a dataset from a scientific discipline A and another from a scientific discipline B. A deepfake science LLM composer could then have learned to surprise a deepfake science LLM 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 have harnessed the knowledge that spontaneous human creativity strongly profits from nocturnal brain processes during sleep (Lewis *et al.*, 2018) to improve the LLM's generation of outcomes perceived to stimulate ideas of implausible utility. To this end, they could have repeatedly fine-tuned the LLM on recursively changing text data modified by loosely mimicking e.g. partially sighted evolutionary affective processes of the spontaneous creativity mode (Aliman, 2020) extending to synergetic cycles of human sleep (Lewis *et al.*, 2018). In simpler cases, this could technically have included, e.g. targeted semantic mutations, syntactic-semantic crossover and a form of semantic noise injection followed by autocorrection at the sentence level. In extensions of such conceptual ideas, scientists could have enriched this shifting dataset by letting the LLM actively integrate scientific knowledge sampled, e.g. from suitable knowledge graphs. Simple active forgetting mechanisms to reduce data size and complexity could have been for instance steered by integrating human preferences via the AU encoding and/or by integrating human attention during interactions with the LLM.

10.3.3 Cyborgnetic defence against LLM-generated reviews

As can be extracted from the last subsection, scientists could have practically transformed the initial merely imitative LLM into an interactive multiversal transdisciplinary deepfake science incubator. The interesting aspect thereby is that this advanced interactive LLM incubator would still not be able to understand and create explanatory knowledge. This signifies that it could have been utilised as a strong baseline offering an enormous amount of material to train the epistemic defences of reviewers against SEA AI attacks. In theory, any conjectured approach to an explanatory IPS to shield peer-review from the non-explanatory contents of SEA AI attacks must be at least robust against the outputs of that LLM incubator at test time. Generally, this could already

have deeply impacted the nature of peer-review. Thereby, the interactive LLM incubator could also have been utilised for autodidactic purposes and to prepare for the red teaming and penetration testing procedures that we already hinted at previously. Strikingly, many of the aforementioned could have led to a human sense of empowerment emerging from the cyborgnets of defenders via the augmentative feedback loops with LLMs. Simultaneously, this could have encouraged an increased awareness of responsibility on the part of the reviewers potentially paired with an altered nature of self-experience via the immensely extended field of affordances for human creativity. This explains why in the quest to defend against SEA AI attacks, humans and language models could indeed become a sort of ethical regulator (Ashby, 2020) of each other. Moreover, it also brings us back to the end of Section 10.2 where we implied that applying our GCD framework to counter SEA AI attacks could at once engender a convergence of moral programming and security research to counter immoral programming.

10.4 Conclusions

In this paper, we performed an in-depth analysis of how to possibly counteract a severe not yet prevalent but technically feasible case of immoral programming: SEA AI attacks, i.e. deepfake science attacks on (applied) science itself. For instance, malicious actors could exploit language AI for future SEA AI attacks instrumental in performing cyber(crime) and information warfare – which requires a thorough assessment of defence methods now and not in hindsight. To this end, we introduced our transdisciplinary GCD framework that can be utilised as a complementary generic scaffold to craft tailored defences. GCD comprises three generic epistemic defences and seven generic so-called cyborgnetic creativity augmentation measures. Focusing on SEA AI attacks with language AI models, we then instantiated this generic scaffold within one exemplary use case, namely large language models such as GPT-3. We then elaborated on how an LLM itself can be employed to defend against SEA AI attacks with LLMs. Thereby, cyborgnetic feedback loops between scientists and LLMs could offer resilience to SEA AI attacks. In addition, they could also transform the language models into interactive multiversal transdisciplinary deepfake science incubators (generating creativity-stimulating but still non-explanatory outcomes) while simultaneously encouraging the multiversal scientists to stay critical and to engage in explanation-anchored, trust-disentangled, and adversarial scientific knowledge co-creation. Whilst implementing such hybrid defence methods against SEA AI attacks, scientists and language models reciprocally become ethical regulators of each other. In short, counteracting immoral programming and moral programming itself converges within the GCD scaffold. In our view, once a rigorous epistemic elucidation is provided to the general public, humanity as a whole may

profit from creativity fostering deepfake incubators via e.g. language model subscriptions that could be available to everyone, such as is the case with access to the internet. Obvious limitations of our framework could be the need to address emerging plagiarism issues (Dehouche, 2021). Overall, GCD-based solutions to 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, we pointed to the daunting SEA AI elephant in the room and proposed a complementary non-exhaustive solution. GCD could provide cognitively diverse incentives for AI safety and for ongoing efforts in moral programming for which Bart Wernaart (2021) recently set forth a future-oriented road map. As we have seen, the international meta-cyborgnet of multiversal scientists is latently capable of building up resilience to SEA AI attacks. In this vein, may the elephant rest in peace.

References

Aliman, N.M., 2020. Hybrid cognitive-affective strategies for AI safety. Doctoral dissertation, Utrecht University, Utrecht, the Netherlands. https://doi.org/10.33540/203

Aliman, N.M., 2021a. Cyborgnetics – The type I vs. type II split. Utrecht, the Netherlands.

Aliman, N.M., 2021b. Explanatory IPS. Available at: https://nadishamarie.jimdo.com/app/download/10829271571/Explanatory_IPS.pdf?t=1632837017.

Aliman, N.M. and Kester, L., 2021. Epistemic defenses against scientific and empirical adversarial AI attacks. In Workshop on Artificial Intelligence Safety 2021 co-located with the 30th International Joint Conference on Artificial Intelligence AISafety@IJCAI 2021.

Aliman, N.M., Kester, L. and Yampolskiy, R., 2021. Transdisciplinary AI observatory–retrospective analyses and future-oriented contradistinctions. Philosophies, 6: 6.

Ashby, M., 2020. Ethical regulators and super-ethical systems. Systems, 8: 53. https://doi.org/10.3390/systems8040053

Ashby, W.R., 1961. An introduction to cybernetics. Chapman & Hall Ltd, London, UK.

Bartoli, A. and Medvet, E., 2020. Exploring the potential of GPT-2 for generating fake reviews of research papers. In Fuzzy Systems and Data Mining. IOS Press, Amsterdam, the Netherlands, pp. 390-396. https://doi.org/10.3233/FAIA200717

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Bunescu, R.C. and Uduehi, O.O., 2019. Learning to surprise: a composer-audience architecture. In ICCC, pp. 41-48.

Dehouche, N., 2021. Plagiarism in the age of massive generative pre-trained transformers (GPT-3). Ethics in Science and Environmental Politics, 21: 17-23. https://doi.org/10.3354/esep00195

Dietrich, A., 2019. Types of creativity. Psychonomic Bulletin & Review, 26: 1-12. https://doi.org/10.3758/s13423-018-1517-7

Eckert, S., Sumner, C. and Krause, T., 2018, 11 August. Inside the fake science factory. Presentation at DEF CON, 26, Las Vegas, NV, USA.

Eleuther AI, 2021. Available at: https://www.eleuther.ai/

Fallis, D., 2020. The epistemic threat of deepfakes. Philosophy & Technology, 1-21. https://doi.org/10.1007/s13347-020-00419-2

Fink, A., Grabner, R.H., Gebauer, D., Reishofer, G., Koschutnig, K. and Ebner, F., 2010. Enhancing creativity by means of cognitive stimulation: evidence from an fMRI study. NeuroImage, 52: 1687-1695. https://doi.org/10.1016/j.neuroimage.2010.05.072

Franceschelli, G. and Musolesi, M., 2021. Creativity and machine learning: a survey. arXiv preprint arXiv:2104.02726.

Frederick, D., 2020. Against the philosophical tide: essays in Popperian critical rationalism. Critias Publishing, Yeovil, UK.

Frederick, D., 2021. Critique of Brian Earp's writing tips for philosophers. Think, 20: 81-87. https://doi.org/10.1017/S1477175621000063

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N. and Presser, S., 2020. The pile: an 800GB dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027. https://arxiv.org/abs/2101.00027

Hartmann, K. and Giles, K., 2020, May. The next generation of cyber-enabled information warfare. In: 2020, IEEE 12th International Conference on Cyber Conflict (CyCon) 1300: 233-250. https://doi.org/10.23919/CyCon49761.2020.9131716

Hutchinson, J.B. and Barrett, L.F., 2019. The power of predictions: an emerging paradigm for psychological research. Current Directions in Psychological Science, 28: 280-291. https://doi.org/10.1177/0963721419831992

Kaloudi, N. and Li, J., 2020. The AI-based cyber threat landscape: a survey. ACM Computing Surveys (CSUR), 53: 1-34. https://doi.org/10.1145/3372823

Kranenbarg, M.W., Holt, T.J. and Van der Ham, J., 2018. Don't shoot the messenger! A criminological and computer science perspective on coordinated vulnerability disclosure. Crime Science, 7: 1-9. https://doi.org/10.1186/s40163-018-0090-8

Lewis, P.A., Knoblich, G. and Poe, G., 2018. How memory replay in sleep boosts creative problem-solving. Trends in cognitive sciences, 22: 491-503. https://doi.org/10.1016/j.tics.2018.03.009

Mitchell, R., Boyle, B., O'Brien, R., Malik, A., Tian, K., Parker, V., Giles, M., Joyce, P. and Chiang, V., 2017. Balancing cognitive diversity and mutual understanding in multidisciplinary teams. Health Care Management Review, 42: 42-52. https://doi.org/10.1097/HMR.0000000000000088

Ozkan, B.Y., Van Lingen, S. and Spruit, M., 2021. The cybersecurity focus area maturity (CYSFAM) model. Journal of Cybersecurity and Privacy, 1: 119-139. https://doi.org/10.3390/jcp1010007

Popper, K.R., 1963. Conjectures and refutations: the growth of scientific knowledge. Routledge, London, UK.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., 2019. Language models are unsupervised multitask learners. OpenAI blog, 1: 9. Available at: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Rajendran, J., Jyothi, V. and Karri, R., 2011, October. Blue team red team approach to hardware trust assessment. In: 2011, IEEE 29th international conference on computer design (ICCD), pp. 285-288. https://doi.org/10.1109/ICCD.2011.6081410

Ranade, P., Piplai, A., Mittal, S., Joshi, A. and Finin, T., 2021. Generating fake cyber threat intelligence using transformer-based models. arXiv preprint arXiv:2102.04351.

Reynolds, A. and Lewis, D., 2017. Teams solve problems faster when they're more cognitively diverse. Harvard Business Review, 30. Available at: https://hbr.org/2017/03/teams-solve-problems-faster-when-theyre-more-cognitively-diverse.

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

Romero, A., 2021. GPT-3 scared you? Meet Wu Dao 2.0: a monster of 1.75 trillion parameters. Available at: https://towardsdatascience.com/gpt-3-scared-you-meet-wu-dao-2-0-a-monster-of-1-75-trillion-parameters-832cd83db484

Tsao, J.Y., Ting, C.L. and Johnson, C.M., 2019. Creative outcome as implausible utility. Review of General Psychology, 23: 279-292. https://doi.org/10.1177/1089268019857929

Van Noorden, R., 2014. Publishers withdraw more than 120 gibberish papers. Nature News. https://doi.org/10.1038/nature.2014.14763

West, J.D. and Bergstrom, C.T., 2021. Misinformation in and about science. Proceedings of the National Academy of Sciences, 118: e1912444117. https://doi.org/10.1073/pnas.1912444117

Wernaart, B., 2021. Developing a roadmap for the moral programming of smart technology. Technology in Society, 64: 101466. https://doi.org/10.1016/j.techsoc.2020.101466

Zhao, B., Zhang, S., Xu, C., Sun, Y. and Deng, C., 2021. Deep fake geography? When geospatial data encounter artificial intelligence. Cartography and Geographic Information Science, 48: 1-15. https://doi.org/10.1080/15230406.2021.1910075