Article

Transdisciplinary AI Observatory—Retrospective Analyses and Future-Oriented Contradistinctions

Nadisha-Marie Aliman [1,*]**, Leon Kester** [2] **and Roman Yampolskiy** [3]

[1] Department of Information and Computing Sciences, Utrecht University, 3584 CC Utrecht, The Netherlands
[2] TNO Netherlands, 2597 AK The Hague, The Netherlands; leon.kester@tno.nl
[3] School of Engineering, University of Louisville, Louisville, KY 40292, USA; roman.yampolskiy@louisville.edu
* Correspondence: nadishamarie.aliman@gmail.com

Abstract: In the last years, artificial intelligence (AI) safety gained international recognition in the light of heterogeneous safety-critical and ethical issues that risk overshadowing the broad beneficial impacts of AI. In this context, the implementation of AI observatory endeavors represents one key research direction. This paper motivates the need for an inherently *transdisciplinary* AI observatory approach integrating diverse retrospective and counterfactual views. We delineate aims and limitations while providing hands-on-advice utilizing *concrete practical examples*. Distinguishing between unintentionally and intentionally triggered AI risks with diverse socio-psycho-technological impacts, we exemplify a *retrospective descriptive analysis* followed by a *retrospective counterfactual risk analysis*. Building on these AI observatory tools, we present near-term transdisciplinary guidelines for AI safety. As further contribution, we discuss *differentiated* and tailored long-term directions through the lens of two disparate modern AI safety paradigms. For simplicity, we refer to these two different paradigms with the terms *artificial stupidity* (AS) and *eternal creativity* (EC) respectively. While both AS and EC acknowledge the need for a hybrid cognitive-affective approach to AI safety and overlap with regard to many short-term considerations, they differ fundamentally in the nature of multiple envisaged long-term solution patterns. By compiling relevant underlying *contradistinctions*, we aim to provide *future-oriented* incentives for constructive dialectics in practical and theoretical AI safety research.

Keywords: AI safety; AI observatory; retrospective counterfactual risk analysis; artificial stupidity; artificial creativity augmentation; cybersecurity; social psychology; HCI

1. Motivation

Lately, the importance of addressing artificial intelligence (AI) safety, AI ethics and AI governance issues has been acknowledged at an international level across diverse AI research subfields [1–6]. From the heterogeneous and steadily growing set of proposed solutions and guidelines to tackle these challenges, one can extract an important recent motif, namely the concept of an *AI observatory* for regulatory and feedback purposes. Notable early practical realizations with diverse focuses include Italian [7], Czech [8], German [9] and OECD -level [10] AI observatory endeavors. Thereby, the Italian AI observatory project targets the public reception of AI technology and the Czech one tackles legal, ethical and regulatory aspects within a participatory and collective framework. The German AI observatory jointly covers technological foresight, administration-related issues, sociotechnical elements and social debates at a supranational and international level. Finally, the OECD AI Policy Observatory *"aims to help policymakers implement the AI Principles"* [10] that have been pre-determined by the OECD and pertain among others to data use and analytical tools. Theoretical and practical recommendations to integrate the retrospective documentation of internationally occurring AI failures have been presented by Yampolskiy [11] and very recently McGregor [12]. In addition, Aliman [13] proposed to complement such *reactive* AI observatory documentation efforts with *transdisciplinary*

and *taxonomy-based* tools as well as *proactive* security activities. In this paper, we build on the approaches of both Yampolskiy and Aliman and elaborate on the necessity of a *transdisciplinary AI observatory* integrating both reactive and proactive retrospective analyses. As reactive analysis, we propose a taxonomy-based *retrospective descriptive analysis* (RDA) which analytically documents factually already instantiated AI risks. As proactive analysis, we propose a taxonomy-based so-called *retrospective counterfactual risk analysis* [14] (RCRA) that inspects plausible peak *downward counterfactuals* [15] of those instantiated AI risks to craft future policies. Downward counterfactuals pertain to *worse* risk instantiations that *could* have plausibly happened in that specific context *but did not*. While an RDA can represent a suitable tool for a qualitative overview of the current AI safety landscape revealing multiple issues to be addressed in the immediate near-term, an RCRA can supplement an RDA by adding breadth, depth and context-sensitivity to these insights with the potential to improve the efficiency of future-oriented regulatory strategies.

The remainder of the paper is organized as follows—in the next Section 2, we first introduce a simple fit-for-purpose AI risk taxonomy as basis for classification within RDAs and RCRAs for AI observatory projects. In Section 3 and in the subsequent Section 4, we elaborate on aims but also limitations of RDA and RCRA while collating *concrete examples* from practice to clarify the proposed descriptive and counterfactual analyses. In Section 5, we exemplify the requirement for *transdisciplinarily* conceived *hybrid cognitive-affective* AI observatory approaches and more generally AI safety frameworks. In Section 5.1, we provide *near-term* guidelines directly linked to the practical factuals and counterfactuals from RDA and RCRA respectively. Hereinafter, we discuss *differentiated* and bifurcated *long-term* directions through the lens of two recent AI safety paradigms: artificial stupidity (AS) and eternal creativity (EC)—succinct concepts which are introduced in Section 5.2. We provide incentives for future constructive dialectics by delineating central distinctive themes in AS and EC which (while overlapping with regard to multiple near-term views) exhibit pertinent differences with respect to long-term AI safety strategies. Thereafter, in Section 6, we briefly comment on data collection methods for RDAs and idea generation processes for RCRAs. Finally, in Section 7, we summarize the introduced ensemble of transdisciplinary and socio-psycho-technological recommendations combining retrospective analyses and future-oriented contradistinctions.

2. Simple AI Risk Taxonomy

For simplicity and means of illustration, we utilize the streamlined AI risk taxonomy displayed in Figure 1 for the classification of practical examples of AI risk instantiations in the RDA and corresponding downward counterfactuals in the RCRA. This simplified taxonomy has been derived from a recent work by Aliman et al. [16]. (Note that the original taxonomy makes a *substrate-independent* difference between two *disjunct* sets of systems: *Type I* systems and *Type II* systems. While the set of *Type II* systems includes all systems that exhibit the ability to *consciously create and understand explanatory knowledge*, Type I systems are by definition all those systems that *do not* exhibit this capability. Obviously, all present-day AI systems are of *Type I* whereas *Type II* AI is up to now *non-existent*. In fact, the only currently known sort of *Type II* systems are human entities. For this reason, the taxonomy we consider here for RDA and RCRA only focuses on the practically-relevant and already instantiated classes of *Type I* AI risks.) Following cybersecurity-oriented approaches to AI safety [11,13,17,18], we do not only classically zoom in on *unintentional* failure modes but also on *intentional* malice exhibited by malevolent actors. This distinction is reflected in the utilized taxonomy by contrasting AI risks brought about by malicious human actors (risk *Ia* and *Ib*) vs. those caused by unintentional failures and events (risks *Ic* and *Id*). Moreover, the taxonomy distinguishes between AI risks forming themselves at the pre-deployment stage (*Ia* and *Ic*) vs. those forming themselves at the post-deployment stage (*Ib* and *Id*).

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

Figure 1. Simplified overview of main Type I artificial intelligence (AI) risks. Modified from Ref. [16].

3. Retrospective *Descriptive* Analysis (RDA)

3.1. Aims and Limitations

To allow for a human-centered AI governance, one requires a dynamic responsive framework that is updatable by design [19] in the light of novel emerging socio-technological [20–22] AI impacts. For this purpose, it has been postulated to combine proactive and reactive mechanisms in AI governance frameworks in order to achieve an effective socio-technological feedback-loop [19]. An RDA can be understood as a reactive AI governance and AI safety mechanism. More precisely, taxonomy-based RDA documentation efforts could facilitate a detailed especially qualitative overview and valuable opportunity for fine-grained monitoring of the AI safety landscape. It could be harnessed to guide regulatory efforts, inform policymakers and raise sensitivity in AI security, law and the general public. Further, an RDA could inform future ethical and security-aware AI design and guide endeavors to build defense mechanisms for AI systems enhancing their robustness and performance.

In addition to the proposed fourfold qualitative distinction via the classification in risks *Ia*, *Ib*, *Ic* and *Id*, one could also introduce a quantitative parameter for intensity ratings [23] such as harm intensity [13]. Given the harm-based nature of human cognitive templates in morality [24,25], a harm parameter could provide a meaningful shortcut to encode the urgency of addressing specific risk instantiations in practice. However, given the simultaneous perceiver-dependency [25,26] of harm perception in morality which is strongly based on dyadic considerations (the degree to which an intentional agent is *perceived* to inflict damage to a vulnerable patient [25]), corresponding assignments may not generalize. Nevertheless, identifying *peaks* of harm intensity above a certain agreed upon threshold (e.g., starting at the level of lethal risks) from an RDA might represent a responsible strategy with less controversial assignments. (Analogously, as further specified in Section 4, it is meaningful to focus on analytically derived above threshold downward counterfactuals as basis for an RCRA.) Extracted RDA peaks can be useful to calibrate regulations where necessary while avoiding superfluous constraints for multiple stakeholders that could hinder freedom and progress in the AI field.

Obviously, the quality of RDA results depends on data collection methods and an RDA may not reveal a comprehensive overall picture. Generally, AI risk instantiations could stay unreported, overlooked by the manual or automated data sampling or even remain unnoticed in certain contexts despite already existing. Finally, it is important to note that an RDA should *not* be understood as means to *predict* the future. As known from Popper, a society cannot predict the contents of its own future knowledge [27]. This *fundamental* unpredictability is directly relevant to understand limitations of an AI observatory—it can only reveal patterns of the past. There is no guarantee of repetitions and for instance completely unknowable novel threats could emerge via future human malevolent creativity in the form of risk instantiations *Ia* and *Ib* or via yet unknown errors leading to future instances *Ic* and *Id*. Instead of conceiving of an RDA as an oracle, we suggest framing it as a valuable preparative but incomplete tool with certain fundamental and further non-fundamental limitations. How an RCRA can be utilized to tackle one restriction of the latter sort is described in Section 4.

3.2. RDA for AI Risk Instantiations Ia and Ib—Examples

To clarify the implementation of a taxonomy-based RDA for an AI observatory, we briefly analytically document a variety of concrete already instantiated AI risks starting with those linked to *intentional* malice (AI risks *Ia* and *Ib*). For risk *Ia*, the current goals of the human entities in the context of many induced events are mostly either *adversarial* goals hold by malicious actors or *research* goals of white hats and AI security researchers. To provide a simple and compact overview for risk *Ia*, we group the space of these different goals in a set of 6 (unquestionably non-exhaustive) main clusters: 5 adversarial clusters and 1 research cluster conflating the research goals. The aim of the research cluster is to demonstrate the feasibility of malicious AI design motivated by diverse adversarial goals across a variety of domains in order to foster safety-awareness. Beyond that, we consider 1 extra emerging risk pattern, namely *automated disconcertion* [28] which we introduce in a few paragraphs.

First, an *adversarial cluster 1* could be described as grouping the use of generative AI for subsequent (cyber-)crime facilitation for example, via impersonation [29–32]. Striking examples for *adversarial cluster 1* include a deep-learning based voice cloning of the CEO of a UK-based company that enabled a fraudster to acquire ca. $243,000 [32] and a scammer that suceeded to cause a transfer of ca. $287,000 with a deepfake video sample impersonation [30]. Second, one can indentify an *adversarial cluster 2* related to defamation, harassment, revenge and sextortion [33] typically employing deepfake techniques such as deep learning based facial replacement to visually place targeted often female individuals in pornographic video settings they never partook [34]. Third, *adversarial cluster 3* comprises the use of AI for misinformation and disinformation purposes [35] including via fake profiles camouflaged with AI-generated synthetic portraits [36]. Fourth, an *adversarial cluster 4* consists in using deepfake methods (as well as recent applications of deepfakes to virtual reality [37]) for a form of non-consensual voyeurism whereby even underage victims are assumed to be affected in some cases [38]. Fifth, *adversarial cluster 5* includes AI-supported espionage [39] (e.g., via AI-generated fake profile pictures on social media platforms [40]), AI-aided intelligence gathering [41] and controversial AI-supported targeted profiling [42].

Moreover, we identify a *research cluster 1* as described. Notably, security researchers provided proof-of-concepts among others related to designing camouflaged undetectable fake samples usable for other crimes (e.g., adversarial deepfakes bypassing deepfake filters [43] which could be misused to conceal unethical illegal material disguised as deepfakes and furthermore undetected AI-generated fake comments i.a. on a federal public comment website [44]). Recent security work also successfully explored advanced deepfake techniques for improved impersonation, spear-phishing and large-scale disinformation [45]. Yampolskiy crafted a proof-of-concept for an AI-generated fake academic article [46] perhaps simultaneously acting as cautionary example and as a form of honeypot [47] for inattentive readers that might cite this article unknowingly. Other researchers identified an emerging interest for deepfake ransomware [48] in certain cybercriminal circles. Beyond that, it has been demonstrated that via a replica of a victim intelligent system (a deep reinforcement learning agent), the policies of the victim system can be compromised in a targeted way [49].

Interestingly, an already perceptible consequence of *the mere existence* of risk *Ia* instantiations containing the design of deepfake technologies already led to the emergence of a risk pattern which has been termed automated disconcertion [28]. Automated disconcertion can imply the intentional or also unintentional mislabelling of real samples as fake—for example, in the context of misleading conspiracy theories [50] or against the background of uncertain political settings as it was the case in Gabon not long ago [51]. (To summarize the latter, a *"recent failed military coup in the context of pre-existing political unrest in Gabon was partially grounded in the proliferation of the wrong assumption that an official presidential video represented a manipulative deepfake video"* [28]). Conversely, automated disconcertion can also mean that fake samples are considered as being authentic or simply lead to highly

uncertain and inconclusive settings in which doubts cannot be further resolved in reasonable time with acceptable resources. In short, this additional outlier risk pattern is called *automated* disconcertion since it does not further necessitate the interference of any actors to be repeatedly instantiated after initiation.

Coming to risk *Ib*, its instantiations are currently predominantly concentrated in a single research-oriented cluster (in analogy to *research cluster 1* for risk *Ia* instantiations). However, it is thinkable that exploits of AI vulnerabilities unknown to the public are already taking place before disclosure (a type of zero-day exploits [52] applied to the AI domain). The main benign research goal for security researchers to target risk *Ib* instantiations is currently mostly to disclose existing AI vulnerabilities against malicious attacks and explore possible novel defenses against those before their exploitation. This already led to an incessant attacker-defender race in the fast moving field of security for machine learning and adversarial examples [53–56]. In recent years, researchers have among others developed different attack schemes on how to evade cybersecurity AI [57], e-mail protection, verification tools [58], forensic classifiers [59] and person detectors [60], how to elicit algorithmic biases [13,61], how to fool medical AI [62–65], law enforcement tools [66] as well as autonomous vehicles [67,68], how to perform denial-of-service and other adversarial attacks on commercial AI services [69–71], how to cause energy-intense and unnecessarily prolonged processing time [72] and how to poison AI systems post-deployment [73].

3.3. RDA for AI Risk Instantiations Ic and Id—Examples

In this subsection, we continue to elucidate the practical application of a taxonomy-based RDA by now briefly analytically documenting various already instantiated *unintentionally* triggered risks that formed themselves at the pre- and post-deployment stage (i.e., risk *Ic* and *Id* respectively). For risk *Ic*, we group the space of observed failure modes in a set of 5 (unquestionably non-exhaustive) main failure clusters. In addition, we present 1 extra emerging risk pattern. In analogy to the outlier risk pattern of automated disconcertion related to risk *Ia* instantiations, we introduce the risk pattern of *automated peer pressure* representing an already perceptible side-effect of specific risk instantiations *Ic*. In the case of AI risk instantiations *Id*, we consider a single main failure cluster. (Overall, in some cases, it is difficult to delineate a risk instantiation type unambiguously (e.g., *Ic* vs. *Id* in the presence of multiple complex influences or even in a few cases *Ic* vs. *Ia* given different ethical perspectives). This practical limitation is partially linked to the perceiver-dependency of classification-related assignments that may also play a role in a future AI observatory. However, by publicly sharing the sources, it is possible for entities external to an AI observatory to refine interpretations. Generally, we humbly subscribe to the epistemological view that all knowledge is fallible [74]).

For risk *Ic*, we consider the 5 main failure clusters described in the following. First, *failure cluster 1* comprises ethically-relevant instances of algorithmic bias [75]. Part of this cluster are misclassifications of diverse underrepresented patterns in AI training datasets with unethical repercussions as exhibited in for example, facial misidentification [76], facial recognition failures [77,78], inaccuracy in AI-aided diagnosis [79]. Other cases are datasets with historically outdated unethical labels [80] and ethically-sensitive training biases favoring overrepresented patterns [81]. Second, *failure cluster 2* refers to instances of poorly designed low-performing AI that are halted subsequently [82]. Third, *failure cluster 3* are AI methods designed for law enforcement but threatening privacy [83]. Fourth, *failure cluster 4* subsumes all unintentional risk instantiations linked to more or less hidden pseudo-scientific or outdated and previously refuted preconceptions. For instance, the deployment of AI for facial recognition of criminals based on *"minute features"* [84,85] in their face is based on pseudo-scientific assumptions [86]. Further, the deployment of present-day image-based "emotion recognition" AI is not grounded in state-of-the-art [87] affective science and lacks the required multimodal and context-sensitive modelling to be able to mimick how humans *infer* [88] (and not detect) affective patterns. In fact, a ban has been

requested for premature emotion AI i.a. to prevent usage in ethically sensitive settings [89] such as law enforcement, fraud detection or recruiting. Fifth, *failure cluster 5* is linked to affective, persuasive [90] and (micro-)targeted AI-aided methods that already permeate human cognitive-affective constructions in a way extending *beyond the initial design purposes* and causing *epistemic* biases ranging from a loss of critical stance via AI-empowered social media [91,92] to flawed mind perception in present-day robots [93,94].

A further risk pattern that emerged via *the mere existence* of specific AI risk instantiations *Ic* assignable to the *failure cluster 5*, is a construct that we call *automated peer pressure*. It is already known that attention at a collective level can be *intentionally* biased and manipulated in social media [91] also with the help of bots [95,96] (risk *Ia*). Moreover, as stated in an open letter written by multiple known psychologists and sent to the American Psychological Association: *"[...] the desire for social acceptance and the fear of social rejection are exploited by psychologists and other behavior change experts to pull users into social media sites and keep them there for long periods of time"* [97]—especially children [90]. Susceptible collective attention mechanisms and beliefs are already even *unintentionally* [92] strongly influenced by AI-empowered social media initially developed for benign purposes. Paired with the strong social dependency of humans where social pressure plays an important regulatory role with biological roots [98], it already triggered what one could call automated peer pressure, a self-perpetuating pattern of social pressure [99–103] without the need for social agents that directly and consciously exert it. Beyond that, the known group phenomenon of *"self-reinforcing networks of like-minded users"* [96] encountered in social media has been termed *homophily* [91,96]. Overall, a combination of a multiplicity of heterogeneous factors of which epistemic biases, homophily, affective contagion [91,104], bots and automated peer pressure are only a subset may foster the documented spread of propaganda in social media [96] as well as the reported negative impacts on the mental health of young users [92,105].

Finally, concerning AI risk *Id*, we observe one main failure cluster which is connected to unanticipated post-deployment usage modes and contexts which also includes eventual complications within unusual interactions of the AI system in a dynamically changing environment. Notable examples are failures of facial recognition AI linked to COVID-19 causing the widespread use of facial masks [106–108], the invariant responses of natural language processing systems when faced with nonsensical instead of usual meaningful queries [109] (disclosing the low level of understanding) and the AI-based censorship of a picture displaying ancient slavery settings due to a forerunning misclassification labelling the sample as displaying nudity [110]. Other cases include unknown latent biases in medical AI [111] and other forms of biases in medical AI that unfold post-deployment as a function of geographical factors [112].

4. Retrospective *Counterfactual* Risk Analysis (RCRA)

4.1. Aims and Limitations

While *upward counterfactuals* of a factual event refer to the better ways in which that event could have unfolded but did not, *downward counterfactuals* refer to those conceivable ways in which this event could have turned out worse. In the past, counterfactual thinking has often been framed as detrimental rumination or even as cognitive bias. However, a modern explanatory framework from social psychology termed *functional theory of counterfactual thinking* (abbreviated with FTCT in the following) stresses that counterfactual thoughts can offer *"[...] insights that comprise blueprints for future action [...]"* [15]. FTCT stresses that counterfactual thinking serves problem-solving and can exhibit high usefulness especially in *complex multi-causal domains* [15]. At the intrapersonal level, counterfactual thoughts are based on implicit processes caused by problems, they are linked to a negatively valenced state of core affect [113] and have the potential to evoke (mental or physical) actions that can potentially correct the underlying errors. This procedure instantiates a regulatory loop—which corresponds to a type of negative feedback model [113] enacted as goal-oriented corrective behavior.

Recently, the notion of an RCRA [14] building upon downward counterfactuals from historical events has been proposed to risk stakeholders in the context of risk management applied to hazardous events (such as earthquakes or terroristic attacks). As explained by Woo [14], such an innovative augmented historical analysis represents a generic universal tool that can supplement regulatory resilience tests and sense-making while facilitating the formation of more differentiated and nuanced views. Given its conjectured *domain-general* nature and seeming applicability to complex multi-causal domains of risk analysis, we suggest to transfer RCRA to AI observatory contexts *at a conceptual level*.

For illustration purposes, the Section 4.3 presents a simplified *RDA-based RCRA* which directly builds upon the exemplary RDA performed in the last Section 3. Our method is *loosely* inspired by Woo's RCRA conception which manifests itself by the general integration of downward counterfactuals from historical samples. However, our step-wise methodology (elucidated in the subsequent Section 4.2) to extract meaningful candidates for the simulation[1] of downward counterfactuals given a large state space of past events has been independently conceived and tailored to the specific AI observatory domain. Overall, we understand an RCRA as complement for a forerunning RDA. Together, this pair of retrospective analyses could provide a solid starting point for future AI observatory projects to be however necessarily updated and error-corrected with time.

In abstract terms, combining RDA and RCRA can be seen as a socio-technological enactment of the regulatory loop-governing behavior [113] described in FTCT—which fits to the AI governance recommendation mentioned earlier in Section 3.1, namely the notion of a socio-technological feedback-loop combining proactive and reactive measures [13,19,20]. While an RDA mainly represents a *reactive* documenting approach, an RCRA *attempts* to broaden future *proactive* measures by anticipating potential extreme branches of the future while resisting the fallacy to cast itself as oracle tool. We emphasize that in the light of the fundamental unpredictability of future knowledge creation as well as the fallibility and incompleteness of human knowledge, surprises and errors are unavoidable. No RCRA can guarantee unassailability. This is similarly the case in cybersecurity for other types of techniques that are likewise assignable to a broad class of proactive security measures related to downward counterfactuals such as penetration testing [114] and red teaming [115,116]. Also there it holds that the non-detection of a vulnerability does not guarantee its absence. (Conversely, the detection of a vulnerability does also not guarantee its future exploitation by malicious actors[2]).

4.2. Preparatory Procedure

After having expounded on aims and limitations of an RDA-based RCRA, we speak to the preparatory procedure of meaningfully extracting the required downward counterfactuals for an RCRA taking as input the set O_{RDA} containing *all instances* from the forerunning RDA. However, before providing further details, we recall as mentioned in Section 3.1 that a meaningful agreed upon threshold τ of harm intensity is recommendable. Although perceiver-dependent, a sufficiently high threshold such as for example, when set to plausible downward counterfactuals of *at least lethal* dimension may be suitable. On an oversimplified harm intensity scale with *1* standing for almost no harm and *5* for existential risk, let *4* stand for a lethal risk (with *2* encoding minor and *3* major harm). Naturally, this threshold and scale are solely employed for purely illustrative purposes and more differentiated and tailored approaches may be required in practice [13]. Equipped with the scale and the exemplary threshold $\tau = 4$, we elaborate in the following on how the set O_{RCRA} of all *clusters*[3] considered in an RCRA can be constructed starting with O_{RDA}

---

[1] Downward counterfactuals can be (co-)created for example, in a predominantly mental form, facilitated by immersive design fiction settings [28] (including storytelling narratives and virtual reality) or simulated and visualized with technological tools.

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

[3] For an enhanced context-sensitivity and to avoid overfitting to the idiosyncrasies of single isolated events, we recommend RCRA simulations at the level of clusters and not of single instances as becomes apparent in the next Section 4.3.

and consecutively applying the following ordered sequence of 4 operations in yet to be described ways: (1) *taxonomization*, (2) *analytical clustering*, (3) *brute-force deliberation and threshold-based pruning*, (4) *assembly*.

As first step, *taxonomization* is applied to O_{RDA} which consists in a one-to-one mapping of each AI risk instantiation sample to either a key from the taxonomy (i.e., *Ia*, *Ib*, *Ic* or *Id*) or in theory to a generic placeholder key for novel unknown patterns. In our description, all samples were directly or at least secondarily assignable to the pre-existing taxonomy keys and no unknown key was required. As second step, the researchers apply an *analytical clustering* operation based on a self-generated explanatory semantic grouping linking every sample associated with a specific key, to a cluster. By way of example, under risk *Ia* discussed in Section 3.2, this operation led to 5 adversarial clusters, 1 research cluster and 1 extra cluster. In a third step, the researchers apply *brute-force deliberation*[4] *and threshold-based pruning* by mentally going through every single sample of O_{RDA} and trying to devise—within reasonable self-determined time limits—a plausible downward counterfactual where it holds for the self-rated harm intensity h, that $h \geq \tau$. If such a suitable downward counterfactual is generated in time, the sample is *maintained*, otherwise the sample is discarded from further consideration. Finally, the fourth operation *assembly* is performed which requires to assemble O_{RCRA} by linking back the remaining samples to their clusters from the second step. On this basis, one obtains the generic downward counterfactuals that need to be analyzed for the intended RCRA. In short, this simple step-wise procedure takes RDA instances as inputs and produces a set of generic RCRA clusters as output. This output set O_{RCRA} represents the superset of the searched meaningful generic above threshold downward counterfactuals of interest. For clarification, the next paragraph briefly comments on the application of this simple preparatory procedure to our exemplary RDA instances.

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

---

[4]　In theory, this search can be optimized further. However, the aim is to (at a later stage) obtain *a broad as possible* set of counterfactual instances to increase illustrative power. Both one-to-one and many-to-one mappings between downward counterfactual instances and clusters can potentially become RCRA-relevant if stored. This is connected to the *complementary cognitive co-creation* method used to interlink the preparatory procedure with the RCRA that we explain in Section 6.2.
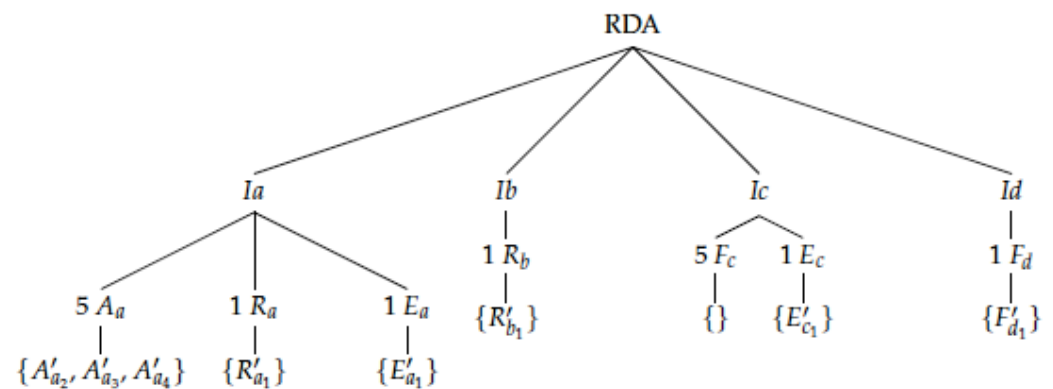
Figure 2. Simplified sketch on possible preparatory procedure to extract peak generic downward counterfactuals for a retrospective counterfactual risk analysis (RCRA) out of a forerunning taxonomy-based retrospective descriptive analysis (RDA) for an AI observatory. The top node stands for the initial set O_{RDA} containing all RDA samples. For illustration, the risk instantiation clusters from Sections 3.2 and 3.3 are filled in. *A* refers to *adversarial*, *R* to *research*, *E* to *extra* and *F* to *failure* cluster. The conjunction of all analytically derived leaves are possible generic above threshold downward counterfactuals of interest for the RCRA. In this example, the output set for the RCRA corresponds to $O_{RCRA} = \{A'_{a_2}, A'_{a_3}, A'_{a_4}, R'_{a_1}, E'_{a_1}, R'_{b_1}, E'_{c_1}, F'_{d_1}\}$. For more details, see text.

4.3. Exemplary RDA-Based RCRA for AI Observatory Projects

Recently, co-creation *design fictions* (DFs) [117,118] known from human-computer interaction (HCI) [119] have been recommended for security practices in the AI field [120] and at the intersection between AI and virtual reality (AIVR) [28]. Generally, DFs *"can be used for technological future projections by experts in the form of for example, narratives or construed prototypes that can be represented in text, audio or video formats but also in VR environments"* [28] (whereby the authors cautioned likewise *not* to regard a DF as a means to *predict* the future but as preparatory tool). In our view, one promising way to perform an RDA-based RCRA could be to frame each RCRA cluster as co-creation DF task. Distinctively, instead of projecting into the future as performed in classical DF contexts, such an *RDA-based RCRA-DF* construes instances of RCRA cluster narratives (or experiential prototypes) projecting to the counterfactual past. For illustration, we apply a simplified RDA-based RCRA-DF to each of the 8 elements within the set $O_{RCRA} = \{A'_{a_2}, A'_{a_3}, A'_{a_4}, R'_{a_1}, E'_{a_1}, R'_{b_1}, E'_{c_1}, F'_{d_1}\}$ assembled in the preparatory procedure of the previous Section 4.2. For RCRA-DFs pertaining to intentional malice (risks *Ia* and *Ib*), we provide short DF narratives taking the form of a succinct *threat model* [28,121] specifying adversarial goals, knowledge and capabilities. By contrast, for unintentional failure modes (risks *Ic* and *Id*), we instead describe a short failure model comprising initial design goals, knowledge gaps and unintended effects. Generally, we only consider instantiations of RCRA clusters that correspond to above threshold downward counterfactuals (i.e., with a harm intensity $h \geq \tau$ whereby in Section 4.2, τ was exemplarily set to lethal risk dimensions).

4.3.1. Downward Counterfactual DF Narrative A'_{a_2}

- *Adversarial Goals:* AI-aided defamation, revenge, harassment and sextortion.
- *Adversarial Knowledge:* Since it is a malicious stakeholder that is designing the AI, the system is available to this adversary in a transparent *white-box* setting. Concerning the knowledge pertaining to the human target, a *grey-box* setting is assumed. Open-source intelligence gathering and social engineering are exemplary tools that the adversary can employ to widen its knowledge of beliefs, preferences and personal traits exhibited by the victim.
- *Adversarial Capabilities:* In the following, we briefly speak to exemplary plausible counterfactuals of at least lethal nature that malicious actors could have been capable

to bring about and that are *"worse than what actually happened"* [14] (as per RDA). For defamation purposes, it would have been for instance possible to craft AI-generated fake samples that wrongly incriminate victims with *not* actually executed actions (e.g., a fake homicide but also fake police violence) leading to a subsequent assassination when deployed in precarious milieus with high criminality. To enact revenge with lethal consequences in socio-cultural settings that particularly penalize the violation of restrictive moral principles, similar AI-based methods could have been applicable (e.g., via deepfakes assumingly displaying fake adultery or contents linked to homosexuality). An already instantiated form of AI-enabled harassment mentioned in the RDA consists in sharing fake AI-generated video samples of pornographic nature via social media channels [34]. Consequences could include suicide of vulnerable targets (as generally in cybervictimization [122]) or exposure to a lynch mob. In fact, the contemplation of suicide by deepfake pornography targets has already been reported lately [33]. Finally, concerning AI-supported sextortion, warnings directed to teenagers and pertaining to the convergence of deepfakes and sextortion have been formulated recently [123]. Given the link between sextortion and suicide associated with motifs such as i.a. hopelessness, humiliation and shame [124], consequences of technically feasible but not yet instantiated deepfake sextortion scams could also include suicide—next to simplifying this criminal enactment by adding automatable elements.

4.3.2. Downward Counterfactual DF Narrative A'_{a_3}

- *Adversarial Goals:* AI-aided misinformation and disinformation.
- *Adversarial Knowledge:* Identical to adversarial knowledge indicated in Section 4.3.1.
- *Adversarial Capabilities:* Technically speaking, a malicious actor could have crafted misleading and disconcerting fake AI-generated material that could be interpreted by extreme endorsers of pre-existing misguided conspiracy theories as providing evidence for their beliefs inciting them to subsequent lethal violence. A historical precedent of gun violence as reaction to fake news seemingly confirming false conspiracy theories was the Pizzagate shooting case where a young man fired a rifle in a pizzeria *"[...] wrongly believing he was saving children trapped in a sex-slave ring"* [125]. Beyond that, when it comes to (micro-)targeted [91] disinformation, conceivable malicious actors could have more systematically already employed hazardous AI-aided information warfare [96] techniques in social media. This could have been supported by AI-enabled psychographic targeting tools [91] and via networks of automated bots [96,126] partially concealed via AI-generated artefacts such as fake profile pictures. While the level of sophistication of many present-day social bots is limited [127], more sophisticated bots emulating a breadth of human online behavior patterns are already developed [128,129] and it is known for some time [130] that *"[...] political bots exacerbate political polarization"* [131]. By AI-aided microtargeting of specific groups of people that are ready to carry out violent acts, malicious actors could have caused more political unrest with major lethal outcomes. In fact, Tim Kendall who was a prior director of monetization at Facebook recently stated more broadly that *"[...] one possible near-term effect of online platforms' manipulative and polarizing nature could be civil war"* [92].

4.3.3. Downward Counterfactual DF Narrative A'_{a_4}

- *Adversarial Goals:* AI-aided non-consensual voyeurism.
- *Adversarial Knowledge:* Identical to adversarial knowledge indicated in Section 4.3.1.
- *Adversarial Capabilities:* Before delving into downward counterfactuals that corresponding malicious actors could have already brought about, it is important to note that the goal considered in this cluster is not primarily the credibility or appearance of authenticity exhibited by the synthetic AI-generated contents. Rather, the focus when visually displaying the target non-consensually in compromising settings is

more on feeding personal fantasies or facilitating a demonstration of power [37,38] while the synthetic samples can obviously concurrently be shared via social media channels. Against this backdrop, it is not difficult to imagine that when editing visual material of vulnerable targets with practices such as deep-learning based "undressing" [38], a disclosure could induce motifs of hopelessness, humiliation and shame in some of those individuals provoking suicidal attempts similar to the hypothetical deepfake sextortion counterfactual described in Section 4.3.1. The mere sensing of having been victimized via non-consensual deepfake pornography has also been associated with the perception of a "digital rape" [33,132]. Especially when the victims are underage [38], this could plausibly reinforce suicidal ideation. Another dangerous avenue may be subtle *combination possibilities* available to the malicious actor. Non-consensual voyeuristic (but also more generally abusive) illegal but quasi-untraceable material bypassing content filters could be meticulously concealed with deepfake technologies and unnoticedly propagated[5] for some time. This could hinder criminal prosecution and particularly threaten the life of vulnerable young victims. Potentiated with automated disconcertion, it could cause a set of latent lethal socio-psycho-technological risks.

4.3.4. Downward Counterfactual DF Narrative R'_{a_1}

- *Adversarial Goals:* Research on malevolent AI.
- *Adversarial Knowledge:* Identical to adversarial knowledge indicated in Section 4.3.1.
- *Adversarial Capabilities:* To begin with, note that in this RCRA cluster, we assume that the research is motivated by *malign* intentions contrary to the corresponding factual RDA research cluster that is conducted with benign and precautionary intentions by security researchers and white hats as mentioned in Section 3.2. This additional distinction is permissible due to its property as *downward* counterfactual. By way of illustration, malicious actors could have already performed research on malevolent AI design in the domain of autonomous mobility or in the military domain. They could have developed a novel type of meta-level physical adversarial attacks on intelligent systems[6] directly utilizing other physically deployed intelligent systems under their control. Such an attacker-controlled intelligent system could be employed as a new advanced form of present-day physical adversarial examples [60,69,134–137] against a selected victim intelligent system. The maliciously crafted AI could have been designed to optimize on physically fooling the victim AI system once deployed in the environment for example, via physical manipulations at the sensor level such as to misleadingly bring about victim policies with lethal consequences entirely unintended by the operators of the victim model. A further concerning instance of malign research could have been secretive or closed-source research on automated medical AI forgery tools that add imperceptible adversarial perturbations to inputs such as to cause tailored customizable misclassifications. While the vulnerability of medical AI to adversarial attacks is already known [62–65,138] and could be exploited by actors intending medical fraud for example, for financial gains, certain exertions of this practice in the wrong settings could be misused as tool for murder attempts and targeted homicides.

---

[5] For instance by mixing real material with synthetic elements obtained from style-based generative adversarial network methods [133], deep-learning based face-replacement and *adversarial deepfake* techniques [43] in order to evade content filters critical to law enforcement.

[6] With intelligent systems, we refer to technically feasible AIs able to independently perform the OODA-loop (i.e., observe, orient, decide, act), but simultaneously totally subordinated to and goal-governed by human entities (e.g., using updatable *human-defined* ethical goal functions [19] prepared pre-deployment—where humans fill in ethically-relevant parameters into a suitable blank but context-sensitive scientifically-grounded template denoted augmented utilitarianism [13]).

4.3.5. Downward Counterfactual DF Narrative E'_{a_1}

- *Adversarial Goals:* This extra cluster of automated disconcertion refers to a risk pattern that emerged automatically from the mere availability and proliferation of deepfake methods in recent years. However, it is conceivable that this AI-related agentless automatic pattern can be intentionally instrumentalized in the service of other (not necessarily AI-related) primary adversarial goals. One example for a primary adversarial goal cluster in the light of which it is appealing for a malicious actor to strategically harness automated disconcertion, would be *information warfare and agitation on social media*. In fact, early cases may already occur [50].
- *Adversarial Knowledge:* Identical to adversarial knowledge indicated in Section 4.3.1.
- *Adversarial Capabilities:* The use of social media in information warfare has been described to be linked to the objective to intentionally blur the lines between fact and fiction [91]. The motif of automated disconcertion itself could be weaponized and misleadingly framed as providing evidence for post-truth narratives offering an ideal breeding ground for global political adversaries performing information warfare via disinformation. Malicious actors could then intensify this framing with the use of pertinent AI technology enlarging their adversarial capabilities as described earlier under the cluster of AI-aided misinformation and disinformation in Section 4.3.2. Given that automated disconcertion may aggravate pre-existing global strategically maintained confusions [139], it becomes clear that a more effective incitement to lethal violence, political unrest with major lethal outcomes or civil wars could be achieved.

4.3.6. Downward Counterfactual DF Narrative R'_{b_1}

- *Adversarial Goals:* Research on vulnerabilities of deployed AI systems.
- *Adversarial Knowledge:* *Grey-box* setting (partial knowledge of AI implementation details).
- *Adversarial Capabilities:* As analogously described in Section 4.3.4, we assume that the research is conducted with malicious intentions. Zero-day exploits of vulnerabilities in (semi-)autonomous mobility and cooperative driving settings to trigger extensive fatal road accidents seem realizable.

4.3.7. Downward Counterfactual DF Narrative E'_{c_1}

- *Designer Goals:* Although automated peer pressure refers to an agentless self-perpetuating mechanism that emerged through AI-empowered (micro-)targeting[7] on social media, its origins can certainly be traced back to the original benign or neutral economic intentions underlying the early design of social media platforms. Psychologist Richard Freed called present-day social media an *"attention economy"* [90] and it is plausible that social media profits from the maximization of utilization time spent by their users.
- *Knowledge Gaps:* Early social media designers may not have foreseen the far-reaching consequences of the designed socio-technological artefacts including threats of lethal dimension or even existential caliber according to some present-day viewpoints [92].
- *Unintended Failures:* The more attention users pay to social media contents, the more time they may spend with like-minded individuals (consistent with

---

[7]   As for instance successfully performed in the Cambridge Analytica case [91].

homophily[8] [91,96]) and the more they may be prone to automated peer pressure. The latter can an also be partially fueled by social bots aggravating polarization [131]. The bigger the success of information warfare and targeted disinformation on social media and the higher the performance of the AI technology empowering it, the more groups of like-minded peers could (but of course not necessarily) uptake misleading ideas. Individuals could then—via these repercussions—sense *a social pressure to suppress their critical thinking* and get accustomed to simply copy in-group narratives irrespective of their contents. This scenario could in turn play into the hands of malicious actors of the type mentioned in Section 4.3.5 and raise the amount and intensity of the lethal and catastrophic scenarios of the sort described in Section 4.3.2.

4.3.8. Downward Counterfactual DF Narrative F'_{d_1}

- *Designer Goals:* Implementation of high-performance AI.
- *Knowledge Gaps:* Designers cannot predict the emergence of yet unknown global risks for which no scientific explanatory framework exists (otherwise that would contradict the fundamental unpredictability of future knowledge creation mentioned in Section 4.1). Given that the past does not contain data patterns of yet never instantiated hazards, the datasets utilized to train "high-performance" AI cannot already have these eventualities reflected in their metrics.
- *Unintended Failures:* Exemplary failures that resulted from this unavoidable type of knowledge gap, are multiple post-COVID AI performance issues [138,153,154]. Simultaneously, humanity relies more and more on medical AI systems. Would humans have been confronted with a more aggressive type of yet unknown biological hazard requiring even faster reactivity, it is conceivable that under the wrong constellations, the AI systems optimizing on metrics pertaining to the then deprecated old or on the novel but yet too scarce and thus biased datasets [154] could have led to unreliable policies up to the potential of a major risk.

5. Discussion

5.1. *Hybrid Cognitive-Affective AI Observatory—Transdisciplinary Integration and Guidelines*

In this Section 5.1, we compile near-term AI safety guidelines with respect to: (1) the factual RDA clusters introduced in Section 3 and (2) the RDA-based RCRA clusters from Section 4.3. For (2), we only specify the necessary supplementary and non-overlapping guidelines to avoid repetitions.

5.1.1. Near-Term Guidelines for Risks *Ia* and *Ib*
RDA

- A_{a_1}: Clearly, for risk *Ia* instances of *adversarial cluster 1* related to the misuse of generative AI to facilitate cybercrimes (e.g., via impersonation within social engineering phone calls), already known security measures regarding identity check are needed as minimum requirement. A standard approach to mitigate dangers of malevolent impersonation [155] is to go beyond something you are (biometric) [156], and to also require something you know (password) [157] and/or something you have (ID card). Generally, an awareness-raising training of users and employees on social engineering methods including the novel combination possibilities emerging from malicious

---

8　Homophily in social media is a multidimensional construct that can refer to attitudes, beliefs, preferences, appearances across a variety of domains. It is by no means limited to the often discussed case of political homophily [140]. For example, empirical social media studies identified weight-based homophily [141], journalistic homophily [142], homophily in rumor sharing [143], higher perceived homophily by users from collectivistic cultures [144], perceived homophily driving consumer purchase intentions [145] and credibility of information [146], homophilic effects in consumer-website relationships [147], homophily as factor for vlogger popularity [148], ideological hashtag homophily in marketing campaigns [149] and even homophily related to music preferences [150]. Apart from that, it is known in social psychology that *"ingroups are seen as more variable than outgroups"* [151] (especially in individualistic cultures). This could arguably strengthen the (wrong) perception of engaging in heterogeneous online spaces. However, some studies actually found social media patterns diverging from homophily [152]. Hence, it is important to further assess the context-sensitive nature of the phenomenon in future work.

generative AI design seems indispensable. In addition, it may be helpful to systematically complement those measures with old-fashioned but potentially effective pre-approved but updatable private arrangements made *offline* which can also employ offline elements for identity check. For instance, the malicious actor may not be able to react appropriately in real-time if presented with a from his perspective semantically unintelligible inspection question making use of offline pre-agreed upon (dynamically updated) linguistically ciphered insider idioms. The induced confusion could consequently help to dismantle the AI-aided impersonation attempt. Having said this, it is important to analyze the attack surface that the availability of voice cloning and even video impersonation with generative AI brings about when instrumentalized for attacks against widespread voice-based or video-based authentication methods.

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

- A_{a_3}: AI-aided misinformation and disinformation represents a highly complex sociopsycho-technological threat landscape that needs to be addressed at multiple levels using multi-layered [158] approaches. For instance, in a recent work addressing the malicious applications of generative AI and corresponding defenses, Boneh et al. [128] provide a list of directly or indirectly concerned actors: *"authors of fake content; authors of applications used to create fake content; owners of platforms that host fake content software; educators who train engineers in sensitive technologies; manufacturers and authors who create platforms and applications for capturing content (e.g., cameras); owners of data repositories used to train generators; unwitting persons depicted in fake content such as images or deepfakes; platforms that host and/or distribute fake content; audiences who encounter fake content; journalists who report on fake content; and so on"*. Crucially, as further specified by the authors, *"a precise threat model capturing the goal and capabilities of actors relevant to the system being analyzed is the first step towards principled defenses"* [128]. In fact, as briefly adumbrated in Section 4.3, the format of the RDA-based RCRA-DFs we proposed for risk *Ia* and *Ib* was purposefully instantiating exactly that—a threat model. Overall, we thus recommend grounding the development of near-term AI safety defenses (as applied to AI-aided disinformation but also more generally) in RDA-based RCRA-DFs that can be once generated potentially retroactively diversified by novel DF narrative instances tailored to the exemplary actors mentioned by Boneh and collaborators. This could broaden the RCRA results and allow for an enhanced targeted development of countermeasures.

- A_{a_4}: For this AI-aided form of non-consensual voyeurism, the measures of an emergency hotline and a specialized haven as mentioned under cluster A_{a_2} are likewise applicable. Legislators need to be informed on psychological consequences especially for underage victims. While cluster A_{a_2} implied the overt public dissemination of compromising material by what minor individuals would be less at risk given the potential repercussions, the purely voyeuristic case can often be *covert* and attracts

motivational profiles that can target minor individuals [38]. In addition, it might be valuable to proactively inform the general public and also adult population groups susceptible to this issue in order to lift the underlying taboos and to mitigate negative psychological impacts. In the long run, instantiations of this cluster are unlikely to be prevented any more than one can prevent someone fantasizing about someone else. Hence, in the age of fake generative AI artefacts with the virtualization of fake acts of heterogeneous nature normally violating physical integrity in the real-world, it might become fundamentally important to re-assess and/or update societal notions intimately linked to virtual, physical and hybrid body perception in a critical and open dialogue.

- A_{a_5}: With regard to AI-aided espionage, companies and public organizations in sensitive domains need to broadly create awareness especially related to the risk of fake accounts with fake but real appearing profile pictures. For instance, since the generator in a generative adversarial network (GAN) [159] is by design imitating features from a given distribution, advanced results of a successful procedure could appear ordinary and more typical—potentially facilitating a psychologically-relevant intrinsic camouflaging effect. In effect, according to a recent study focused on the human perception of GAN pictures displaying faces of fake individuals that do not exist, *"[...] GAN faces were more likely to be perceived as real than Real faces"*[9] [160]. Beyond that, the authors described an increased social conformity towards faces perceived as real independently of their actual realness. This is concerning also in the light of the extra cluster E_{c_1} of automated peer pressure that could make AI-aided espionage easier. A generic trivial but often underestimated guideline that may also apply to AI-aided open-source intelligence gathering would be to reduce the sharing of valuable information assets via social media channels and more generally on publically available sources to a minimum. Finally, to confuse person-tracking algorithms and prevent AI-aided surveillance misused for espionage, camouflage [161] and adversarial patches [60] embedded in clothes and accessoires can be utilized.

- R_{a_1}: As deep-fake technology proliferates and is used in numerous criminal domains, it is conceivable that an arms-race between malevolent fakers and AI forensic experts [162,163] will ensue, with no permanent winner. Given that this cluster R_{a_1} covers a wide variety of research domains in which security researchers and white hats attempt to preemptively emulate malicious AI design activities to foster safety awareness, a consequential recommendation appears to actively support such research at multiple scales of governance. Talent in this adversarial field would need to be attracted by tailored incentives and should not be limited to a standard sampling from average sought-after skill profiles in companies, universities and public organizations of high social reputation. This may also help to avoid an undesirable drift to adversaries for instance at the level of information operations risking reinforcing capacities mentioned in the downward counterfactual DF narratives on cluster A'_{a_3}, R'_{a_1} and E'_{a_1} presented in Section 4.3. Hence, a monolithic approach in AI governance with a narrow focus on ethics and unintentional ethical failures is insufficient [13]. Finally, we briefly address guidelines related to a specific R_{a_1} issue concerning *science* (as asset of invaluable importance for a democratic society [164]) that did not yet gain attention in AI safety and AI governance but that makes further inspections appear imperative in the near-term. Namely, targeted studies on *AI-aided deception in science* to produce AI-generated text disseminated as *fake research articles* (see the research prototype developed by Yampolskiy [46] in another research context) and possibly AI-generated audiovisual or other material meant to display *fake experiments* or also *fake historical samples* (see the recent MIT deepfake demonstration [165] developed for educative

---

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

purposes). However, this technical research direction requires a supplementation by transdisciplinary experts addressing the socio-psycho-technological impacts and particularly the *epistemic* impacts of corresponding future risk instantiations. We suggest that for a safety-relevant sense-making, AI governance may even need to stimulate debates and exchanges on the very epistemological grounding of science—*before* for example, future texts written by maliciously designed sophisticated AI bots (also called sophisbots [128]) infiltrate the scientific enterprise with submissions that go undetected. For instance, there is a *fundamental discrepancy*[10] between how Bayesian and empiricist epistemology would analyze this risk vs. how Popperian critical rationalist epistemology would view the same risk. Disentangling this epistemic issue is of high importance for AI safety and beyond as becomes apparent in the guidelines linked to the next cluster E_{a_1} below.

- E_{a_1}: Near-term guidelines to directly tackle this extra cluster associated to automated disconcertion seem daunting to formulate. However, as a first small step, one could focus on how to avoid exacerbating it. One reason why this cluster may seem difficult to address is due to its deep and far-reaching *epistemic* implications pertaining to the nature of falsification, verification, fakery and (hyper-)reality [169] itself. With regard to this feature of epistemic relevance, E_{a_1} exhibits a commonality with the just introduced different risk of AI-aided deception in science. We postulate that in the light of pre-existing fragile circumstances in the scientific enterprise including the emergence of modern "fake science" [170] patterns but also the mentioned fundamental discrepancies across epistemically-relevant scientific stances, AI-aided deception in science could have direct repercussions on automated disconcertion. First, it could for instance unnecessarily aggravate automated disconcertion phenomena in the general public as for example, the belief in *epistemic threats* [166] could increase people's subjective uncertainty. Second, a reinforced automated disconcertion can subsequently be weaponized and instrumentalized by malicious actors with lethal consequences as generally depicted under the downward counterfactual DF narrative E'_{a_1} described in Section 4.3. This explains our near-term AI governance recommendation to address AI-aided deception in science as transdisciplinary collaborative endeavor analyzing socio-psycho-technological and epistemic impacts.

- R_{b_1}: For this cluster linked to risk *Ib* and pertaining to research on AI vulnerabilities currently performed by security researchers and white hats, we recommend (as analogously already explained in R_{a_1}) to recruit such researchers preemptively. In this vein, Aliman [13] proposes to *"organize a digital security playground where "AI white hats" engage in adversarial attacks against AI architectures and share their findings in an open-source manner"*. For the specific domain of intelligent systems, it is advisable to proactively equip these AIs with technical self-assessment and self-management capabilities[11] [20] allowing for better real-time adaptability for the eventuality of attack scenarios known from past incidents or proof-of-concept use cases studied by security researchers and white hats. However, it is important to keep in mind that challenges from this cluster also deal with zero-day AI exploits, they are the unknown unknowns and cannot be meaningfully anticipated and prevented, though it is realized that many issues could be caused by under-specification in machine learning systems [171].

---

10   Bayesian and empiricist epistemological stances placing the *empirical collection of evidence* and the identification of *true beliefs* at the center of science may link AI-aided deception to *"epistemic threats"* [166]—knowledge-relevant impairments of belief-updating which they already see emerging via deepfakes (i.a. subsuming a general decrease of information in audiovisual samples [166]). By contrast, Popperian epistemic views [167] and especially their Deutschian extension [168] predominantly emphasize in the first place the *explanatory* and *criticism-centered* purpose of science next to the (experimental) *falsifiability* of hypotheses. Strikingly, Deutsch describes science as the endless quest for invariant, *hard-to-vary explanations* of reality [168]. On this view, AI-aided deception in science may be practically problematic, but without question solvable. In fact, while the empiricist direction faces epistemic threats and a post-truth difficulty, the Popperian and Deutschian direction may neither see explanatory knowledge, truth, falsifiability nor the scientific method per se at risk.

11   The conjunction of technical self-assessment and self-management has been summarized under the synonymous umbrella terms of *Type I AI self-awareness* [13], *self-awareness functionality* [20] or simply *self-awareness*.

RCRA (Additional Non-Overlapping Guidelines)

- A'_{a_2}: Generally, one possible way to systematically reflect upon defense methods for specific RCRA instances (generated from *downward* counterfactual clusters) of harm intensity $h_{down} \geq \tau$, could be to perform corresponding *upward* counterfactual deliberations targeting a harm intensity $h_{up} < \tau$. As briefly introduced in Section 4.1, upward counterfactuals refer to those ways in which a certain event could have turned out better but did not. Recently, Oughton et al. [172] applied a combination of downward and upward counterfactual stochastic risk analyis to a cyber-physical attack on electricity infrastructure. In short, the difference to the method that we propose is that instead of focusing on *slightly better upward counterfactuals given the factual event* as made sense in the case of Oughton et al., we suggest a threshold-based selection of *below threshold upward counterfactuals given above threshold downward counterfactuals*[12]. For instance, as applied to the present downward counterfactual cluster A'_{a_2} which also included a narrative instance describing suicide attempts with lethal outcomes as a consequence of AI-aided defamation, harassment and revenge, it could simply consist in *deliberations on how to avoid these lethal scenarios*. This could be implemented by deliberating from the perspective of planning a human, hybrid or fully automated AI-based emergency team response with a highly restricted timeframe (e.g., to counteract the domino-effect initiated by the deployment of the deepfake sample on social media). Next to a proactive combination of deepfake detectors and content detectors for blocking purposes that can fail, a reactive automated social network graph analysis AI combined with sentiment analysis tools could be trained to detect large harassment and defamation patterns that if paired with the sharing of audiovisual samples, can prompt a human operator. This individual could then decide to call in social services that in turn proactively contact the target offering support as analogously mentioned under the guidelines for the factual RDA sample A_{a_2}.

- A'_{a_3}: For this downward counterfactual cluster on AI-aided misinformation and disinformation of at least lethal dimensions, we focus on recommendations pertaining to journalism-relevant defenses and bots on social media. Disinformation from fake sources could be counteracted with the use of blockchain-based reputation systems [173] to assess the quality of information sources. Journalists could also entertain a collective blockchain-based repository containing all news-relevant audiovisual deepfake samples whose authenticity has been refuted so far. This tool could be utilized as publically available high-level filter to evade certain techniques of disinformation campaigns. Moreover, the case of hazardous large-scale disinformation supported by sophisticated automated social bots is of high relevance for what one can term *social media AI safety*. Ideally, *tests* for a "bot shield" enabling some bot-free social media spaces could be crafted. However, it is conceivable that at a certain point, AI-based bot detection [174] might become futile. Also, social bots already fool people [131,175] and many assume that humans will become unable to discern them in the long-term. Nevertheless, it could be worthwhile viewing what one could have done better already with present technological tools (the upward counterfactuals)—which can also include the consideration of divergent unconventional solutions or novel formulations of questions. As stated by Barrett, *"[...] progress in science is often not answering old questions but asking better ones"* [176]. Perhaps, in the future, humans could still devise bot shielding tactics that could attempt to bypass epistemic issues [177] intrinsic to imitation game and Turing Test [178] derivatives where "real" and "fake" become relative.

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

---

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

(which also included the creation of public awareness and the removal of associated taboos). Social services and public institutions like universities and schools could offer emergency psychological interventions for the person at risk. Next to necessitated measures at the level of legal frameworks to protect underage victims, the subtle case of adult targets calls for instance for a civil reporting office collaborating with social media platforms which could initiate a critical dialogue with the other party to bring about an immediate deletion or at least categorical refraining from further dissemination of the material which can be calibrated to the expectations of the target. Recently, the malicious design of deepfakes has been described as a *"[...] serious threat to psychological security"* [179]. Adult targets may despite the synthetic nature of the deepfake samples and often eventually their private character restricted to a personal possession of the agent in question, perceive their mere existence as degradation [180]—a phenomenon certainly requiring social discourses in the long-term. For a principled analytical approach, an extensive psychological research program integrating a collaboration with i.a. AI security researchers could be helpful in order to be able to contextualize relevant socio-psycho-technological aspects against the background of advanced technical feasibility. Importantly, instead of limiting this research to deepfake artefacts in the AI field, one needs to also cover novel hybrid combination possibilities available for the design of non-consensual voyeuristic material. Notably, this includes blended applications at the intersection of AI and virtual reality [28,37] (or augmented reality [181]).

- R'_{a_1}: Concerning proactive measures against future research where an adversary designs self-owned intelligent systems to trigger lethal accidents on victim intelligent systems, one might require legal norms setting minimum requirements on the techniques employed for the cybernetic control of systems deployed in public space. From an adversarial AI perspective, this could include the obligation to integrate regular updates on AI-related security patches developed in collaboration with AI security researchers and white hats that also study advanced physical adversarial attacks. This becomes particularly important as many stakeholders are currently unprepared in this regard [182]. As guideline, we propose that future adversarial AI research endeavors explore attack scenarios where adversarial examples on physically deployed intelligent systems are delivered by another physically deployed intelligent system which potentially offers more degrees of freedom to the malicious actor. From a systems engineering perspective, any intelligent system might need to at least integrate multiple types of sensors and check for inconsistencies at the symbolic level. Next to explainability requirements, a further valuable feature to create accountability in the case of accidents could be a type of self-auditing via self-assessment and self-management [20] allowing for a retrospective counterfactual analysis on what went wrong.

- E'_{a_1}: As its factual counterpart E_{a_1}, this counterfactual cluster E'_{a_1} refering to automated disconcertion instrumentalized for AI-aided information warfare and agitation on social media with the risk to incite lethal violence at large scales, represents a weighty challenge of international extent. As for E_{a_1}, multi-level piecemeal tactics of constructive small steps such as for example, targeted methods to avoid exacerbating it may be valuable. Concerning AI governance, that could include the strategies mentioned under E_{a_1} but also more general efforts in line with international frameworks that aim to foster strong institutions and error-correction via life-long learning (see e.g., [19] for an in-depth discussion).

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

5.1.2. Near-Term Guidelines for Risks *Ic* and *Id*

As can already be realized from the scope of the AI safety guidelines proposed in Section 5.1.1 which are grounded in our AI observatory exemplification of RDA and RCRA, modern AI technology cannot be analyzed in isolation. In our view, due to the complex multi-causal socio-psycho-technological interwovenness underlying AI risks and their instantiations, AI safety requires an inherently *transdisciplinary*, *hybrid* and *cognitive-affective* approach [13]. Transdisciplinarity is especially required to avoid cognitive blind spots within AI safety risk analyses and formulations of countermeasures or guidelines. AI safety needs a hybrid perspective to incorporate the intricacies of human-computer interactions necessitating a consideration of *human nature* next to purely technological viewpoints. Finally, a cognitive-affective perspective is called for due to the inseparably affective nature of human cognition [183,184] whose disregard in AI development can consequently engender significant safety issues by virtue of a lack of requisite variety [185]. While the last Section 5.1.1 focused on guidelines concerning the AI risks *Ia* and *Ib* related to intentional malice, this Section 5.1.2 is linked to the risks *Ic* and *Id* related to mistakes and unintentional failures which are often of ethically-relevant nature. This specific avenue of research represents a well-studied field at the core of modern AI ethics which recognizes multidisciplinarity, human-centeredness and socio-technical contextualization as important requirements [186]. In the last years, a large multiplicity of heterogeneous AI ethics guidelines have been proposed at an international level [187–190]. We refer the reader to Jobin et al. [191] for a global overview of internationally proposed AI ethics guidelines which are directly of relevance for the 5 failure clusters (F_{c_1} to F_{c_5}) linked to risk *Ic* from the RDA presented in Section 3.3. In the following, we focus on the few remaining RDA and RCRA clusters which are not classically in the primary focus of AI ethics.

RDA

- E_{c_1}: This cluster related to automated peer pressure can be i.a. met by measures raising public awareness on the dangers of the confirmation bias [192,193] reinforced via AI-empowered social media. However, a possible upward counterfactual on that issue would be to revert negative consequences of automated peer pressure by utilizing it for beneficial purposes. For instance, it is cogitable that automated peer pressure need not represent a threat would it simply perhaps paradoxically socially *reinforce critical thinking* instead of reinforcing tendencies to blindly copy in-group narratives. Ideally, such a peer pressure would reinforce *heterophily* (the antonym of homophily) with regard to various preferences with one notable exception being the critical thinking mode itself. Hence, one interesting future-oriented solution for AI governance may be education and life-long learning [19] conveying critical thinking and criticism as invaluable tools for youth and general public. For instance, critical thinking skills fostered in the Finnish public education system were effective against disinformation operations [91]. In fact, critical thinking, criticism and transformative contrariness may not only represent a strong shield to tackle disinformation or automated disconcertion and its risk potentials (cluster E_{a_1} and E'_{a_1} respectively), but it also represents a crucial momentum for human creativity [13,194]. Generally, peer pressure is in itself a psychological tool that could be systematically used for good, for example by creating an artificial crowd [195] of peers with all members interested in desirable behaviors such as education, start-ups or effective altruism. A benevolent crowd of peers can then counteract hazardous bubbles on social media.
- F_{d_1}: Concerning AI failures rooted in *unanticipated* and yet unknown post-deployment scenarios, it becomes clear that accuracy and other AI performance measures cannot be understood as conclusive and engraved in stone. A possible proactive measure against post-deployment instantiations of yet unknown AI risks could be the establishment of a generic corrective mechanism. Problems which AI systems experience during its deployment due to differences between training and usage environments can be reduced via increased testing and continued updating and learning stages.

On the whole, multiphase deployment, similar to vaccine approval phases, can reduce an overall negative impact on society and increase reliability. Finally, for each safety-critical domain in which AI predictions are involved in the decision-making procedure, one could—irrespective of present-day AI performance—foresee the proactive planification of a human response team in case of sudden expanding anomalies that a sensitized and safety-aware human operator could detect.

RCRA (Additional Non-Overlapping Guidelines)

- E'_{c_1}: A twofold guideline for this counterfactual cluster (refering to automated peer pressure with lethal consequences via automated disconcertion instrumentalized for AI-aided disinformation), could be to weaken the influence of social bots by measures described under cluster A'_{a_3} and by transforming automated peer pressure into strong incentives for critical thinking as stated in E_{c_1}.
- F'_{d_1}: Finally, for this cluster of major risk dimension being the counterfactual counterpart of cluster F_{d_1}, we emphasize the importance of an early proactive response team formation in contexts such as for instance medical AI, AI in the financial market, AI-aided cybersecurity and critical intelligent cyberphysical assets. In short, AI systems should by no means be understood to be able to truly operate independently in a given task even if current excellent performance measures seem to suggest so. In the face of unknown and unknowable changes, performance is a moving target which if mistaken as conclusive and static could endanger human lives.

5.2. Long-Term Directions and Future-Oriented Contradistinctions

After having introduced a broad variety of near-term guidelines for future AI observatory endeavors based on the exemplified systematic factual and counterfactual retrospective analyses, we provide a differentiated more *general* outlook on explicitly *long-term* AI safety directions. For this purpose, we select two recent theoretical AI safety paradigms: on the one hand a direction that has been termed *artificial stupidity* (AS) (see [196–198]) and on the other hand, a direction that we succinctly call *eternal creativity* (EC) stemming from recent work [13,16,199]. Thereby, note that these two paradigms are by no means postulated to represent the full panoply of nuances and views across the entirety of the young AI safety field. Rather, we select these specific two examples because critical contradistinctions ascertainable via a comparative analysis point to a set of decisive bifurcations which might be of particular interest for the AI safety community due to their potentially *axiomatic* relevance for the future of AI research. While AS and EC coincide in multiple short-term considerations given their common *hybrid cognitive-affective* nature and their emphasis on *cybersecurity-oriented* practices, they fundamentally differ with regard to 3 future-relevant contradistinctions.

We consider the following 3 contradistinctive leitmotifs: (1) *regulatory distinction criterium*, (2) *regulatory enactment* and (3) *substrate management*. First, while AS primarily considers *intelligence* levels for (1), EC ponders the ability to consciously create and understand *explanatory knowledge*. Second, whilst AS foresees deliberate *restrictions* of AI capabilities as tool for (2), EC especially tackles their systematic *enhancement*. Third, while AS views *substrate-dependent* hardware analyses (next to software considerations) for bounded equalization between humans and AIs as approach to (3), EC aims at unbounded *substrate-independent* functional augmentation. While there exist certainly more possible lines along which one could compare AS and EC, we focus on the mentioned 3 themes due to their urgency and potential to foster constructive dialectics in future theoretical and applied AI (safety) research beyond AI observatory contexts. In Sections 5.2.1 and 5.2.2, we briefly provide a general introduction followed by a summarization of long-term AI safety guidelines formulated from the perspective of AS and EC respectively as seen through the lens of these 3 contradistinctions.

5.2.1. Paradigm Artificial Stupidity (AS)

One core assumption in the AS paradigm is that an artificial general intelligence (AGI) *"[...] can be made safer by limiting its computing power and memory, or by introducing Artificial Stupidity on certain tasks"* [197]. Thereby, an AI system is understood to be made *artificially stupid* on a certain task if its capabilities are deliberately limited by human designers for the purpose of matching the human performance on that task. One mentioned exemplary domain where such a technique is already applied is in text-to-speech synthesis such as e.g in Google Duplex, an AI for natural conversations over the phone whose implementation included *"[...] the incorporation of speech disfluencies (e.g., "hmm"s and "uh"s)"* [200]. Another example is the context of video games where AI can in principle vastly exceed human performance which is however purposefully restricted in order to allow for a positive human-centered gaming experience. More generally, there are many AI application domains where it is human-desirable to mimic anthropic performance or behavioral patterns for an improved customer service. These cases correspond to a type of imitation game which only succeeds if the AI does not reveal latent super-human capabilities. From that point of view, the AS paradigm conceives of making an AI artificially stupid as being necessary to making it pass a Turing test [197,198].

Simultaneously, in the last years, AI achieved superhuman-level performance across more and more tasks. Further, it is assumed in AS that *"[...] AI tends to quickly achieve super-human level of performance after having achieved human-level performance"* [198]. Against this background, AS argues distinguishingly that *"[...] by limiting an AI's ability to achieve a task, to better match humans' ability, an AI can be made safer, in the sense that its capabilities will not exceed humans' capabilities by several orders of magnitude"* [198]. In short, AS postulates that *AI ability needs to be upper-bounded by human performance* since it risks to otherwise become *uncontrollable*[13]—once it turns into what Bostrom termed a *superintelligence*—an intellect exceeding human cognitive performance across *"[...] virtually all domains of interest"* [203]. Such a hypothetical future artificial superintelligence is believed to not necessarily be value-aligned with humans (while potentially becoming unintelligible to humans due to the gaps in performance), to be capable of insidious betrayal (a scenario termed *treacherous turn* [203]) and to potentially represent a major risk [204] to humanity.

- *Regulatory distinction criterium:* In this light, one can extract *intelligence* (or more broadly "performance" or "cogntive performance" across tasks) as the recurring theme of relevance for regulatory AI safety considerations under the AS paradigm. At a first level, one could identify two main safety-relevant clusters: a cluster comprising all AIs that are less or equally capable than an average human [198] and another cluster of superintelligent AI systems. The latter can be further subdivided into three classes of systems as introduced by Bostrom [203]: (1) speed superintelligence, (2) collective superintelligence and (3) quality superintelligence. According to Bostrom, the first ones *"can do all that a human intellect can do, but much faster"*, the second ones are *"composed of a large number of smaller intellects such that the system's overall performance across many very general domains vastly outstrips that of any current cognitive system"* and the third ones are *"at least as fast as a human mind and vastly qualitatively smarter"* [203].
- *Regulatory enactment:* In a nutshell, AS recommends *limiting* an AI in hardware and software such that it does not attain any of these enumerated sorts of superintelligence since *"[...] humans could lose control over the AI"* [197]. AS foresees regulatory strategies on *"how to constrain an AGI to be less capable than an average person, or equally capable, while still exhibiting general intelligence"* [198].
- *Substrate management:* To limit AI abilities while maintaining functionality, AS proposes multiple practical measures at the hardware and software level. Concerning the former it proposes diverse restrictions especially pertaining to memory, processing, clock speed and computing [198]. With regard to software, it foresees necessary limits

---

[13] For an in-depth discussion related to AI uncontrollability and unpredictability, see especially [201] and [202] respectively.

on self-improvement as well as measures to avoid treacherous turn scenarios [197]. Another guideline consists in deliberately incorporating known human cognitive biases in the AI system. More precisely, AS postulates that human biases *"can limit the AGI's intelligence and make the AGI fundamentally safer by avoiding behaviors that might harm humans"* [197]. Overall, the substrate management in AS can be categorized as *substrate-dependent* because the artificial substrate is among others specifically tuned to match hardware properties of the human substrate for at most equalization purposes. In summary, AS suscribes to the viewpoint that AI safety aims to *"limit aspects of memory, processing, and speed in ways that align with human capabilities and/or prioritize human welfare, cooperative behavior, and service to humans"* [196] given that AGI *"[...] presents a risk to humanity"* [196].

5.2.2. Paradigm Eternal Creativity (EC)

According to Deutsch, *"the only uniquely significant thing about humans [...] is our ability to create new explanations [...]"* [168]. He further specifies that explanatory knowledge *"gives people a power to transform nature which is ultimately not limited by parochial factors, as all other adaptations are, but only by universal laws"* [168]. Instead of emphasizing levels of intelligence or of performance across a wide set of tasks when analyzing AI safety issues, EC focuses epistemologically on one unique "task": *the ability to consciously create and understand explanatory knowledge*. Thereby, in EC, explanatory knowledge creation also implies the capability to *consciously* understand. Given that core affect is understood as a fundamental property of consciousness [176,183] and is linked to cognitive-affective counterfactual deliberations [16], this excludes philosophical zombie themes [205]. (In modern embodied and enactive cognition frameworks [176,206], consciousness is linked to processes of inference for the cybernetic control of a substrate in an environment connected to allostasis [184] (anticipation of needs before they occur [176])—integrating predictions and error signals from external and internal milieu. It is on such cybernetic control grounds that affective dynamics give rise to the egocentric virtual first-person perspective of the world [207,208] familiar to humans and lacking in present-day AI).

Note that EC's focus on consciously creating and understanding explanatory knowledge is by no means an anthropomorphic assumption forced on AI systems. As elucidated in constructor theory [209,210], a novel explanatory framework in physics, explanatory knowledge creators (of which currently only humans are known) are brought to the fore in physics in an entirely non-anthropocentric way. To put it very simply, constructor theory focuses on *possible* vs. *impossible* counterfactuals that is, what *could* happen given physical laws and *why* (instead of predictions based on laws of motion and initial conditions). On contemplating the set of all physical transformations that would be *possible* in the universe that is, those that *could* happen, one would notice that the size of the very subset containing those transformations that *actually happen* can be strongly influenced by entities able to create and understand knowledge on how to bring them about [168]. This is how explanatory knowledge creation enters *"the cosmic scheme of things"* [168] and this is also why EC prioritizes the conscious understanding and creation of explanatory knowledge via creativity[14] instead of intelligence.

At first sight, given the fundamental unpredictability of future explanatory knowledge, it might seem dangerous for AI safety. Deutsch mentions that *"no good explanation can predict the outcome, or the probability of an outcome, of a phenomenon whose course is going to be significantly affected by the creation of new knowledge"* [168] and further that this fundamental limitation is something that *"when planning for the future, it is vital to come to terms with*

---

[14]　From a psychological and neurocognitive perspective, EC currently views creativity as a tri-partite evolutionary affective construct with varying degrees of *sightedness* [199] instead of a *blind* evolutionary process without a goal akin to biological evolution—as mistakenly assumed by Popper [16,211]. This is epistemologically relevant because ideas are *not* created by blind trial and error (as variation and selection in biological evolution). Even if novel idea contents are fundamentally unpredictable a priori, idea variation is partially guided by previous experience, the task and contextual cues that is, there is a non-zero coupling between variation and selection [211]. Creativity itself could have historical roots in *serendipity* and multi-purpose socially *shared doubt* [13] facilitating in theory error-correction but initially largely used to maintain traditions.

it" [168]. EC agrees. EC recently formulated the *AI safety paradox* [13,16] stating that value alignment and control are conjugate requirements in AI safety. This means that both prevailing ideals cannot be simultaneously fulfilled. EC also states that *"the price of security is eternal creativity"* [13]. So despite the AI safety paradox, a cybersecurity-oriented and risk-centered AI safety is possible—when reframed *"as a discipline which proactively addresses AI risks and reactively responds to occurring instantiations of AI risks"* [13]. In short, AI safety is not condemned, it just needs to come to terms with the compulsion to keep correcting and creating solutions "ad infinitum".

- *Regulatory distinction criterium:* EC distinguishes two *substrate-independent* and disjunct sets of systems: Type I and Type II systems. Type II systems are all systems for which it is possible to consciously create and understand explanatory knowledge. Type I systems are all systems for which this is an impossible task [15]. Thereby, a subset of Type I systems can be conscious (such as non-human mammals) and requires protection akin to animal rights. Obviously, *all present-day AI systems are of Type I* and *non-conscious*. Type II AI is non-existent today.

- *Regulatory enactment:* In theory, with a Type II AI, *"a mutual value alignment might be achievable via a co-construction of novel values, however, at the cost of its predictability"* [13]. As with all Type II systems (including humans), the future contents of the knowledge they will create are fundamentally unpredictable—irrespective of any intelligence class[16]. In EC, this signifies that: (1) *Type II AI is uncontrollable[17] and requires rights on a par with humans*, (2) *Type II AI could engage in a mutual bi-directional value alignment with humans—if it decides so* and (3) *it would be unethical to enslave Type II AI*. (Finally, banning Type II AI is a potential loss of requisite variety and does not hinder malicious actors to do so.) By contrast, regarding Type I AI, EC implies that: (4) *Type I AI is controllable*, (5) *Type I AI cannot be fully value-aligned* across all domains of interest for humans due to an insufficient understanding of human morality, (6) *conscious Type I AI is possible and would require animal-like rights* but it is clearly *non-existent* nowadays.

- *Substrate management:* To avoid functional biases [214] due to a lack of diversity in information processing, EC opts for a *substrate-independent* functional view. Irrespective of its specific substrate composition, an overall panoply of systems is viewed as *one* unit with diverse functions. Given Type-II-system-defined cognitive-affective goal settings, a systematic function integration can yield complementary synergies. Notably, EC recommends research on substrate-independent functional *artificial creativity augmentation* [199] (artificially augmenting *human* creativity *and* augmenting *artificial* creativity). For instance, active inference could technically *increase* Type I AI exploratory abilities [215,216]. Besides that, in Section 6.2, we apply a functional viewpoint to *augment* RCRA DF generation by human Type II systems for AI observatory purposes.

---

[15] EC could be stated to apply a constructor-theoretic distinction to AI safety insofar as it applies a possibility-impossibility dichotomy [209] embedded in an explanatory framework to it.

[16] Under EC, superintelligence is as explained not of distinctive interest. It is also viewed as *not* implying profound *qualitative* differences to human baselines. Following Deutsch, it would be *"[...] subject only to limitations of speed or memory capacity, both of which can be equalized by technology"* [212]. EC views human augmentation as valid transformative defense strategy [199].

[17] Importantly, note that Type II AI uncontrollability does *by no means* imply that a Type II AI is necessarily more dangerous than an arbitrarily designed Type I AI. First, it is important to consider that already an advanced *Type I* AI could lead to existential risks for instance when maliciously designed by malevolent human actors to operate *"at a global scale (e.g., affecting global ecological aspects or the financial system)"* [16]. Second, while it is obvious that a Type II AI *could* be highly dangerous, this also holds for humans including adult terrorists threatening international safety. Overall, it seems a prejudice to assume that Type II AIs that would be members of an open society would *inherently* tend to opt for immutable goals of indifference or extreme violence (see e.g., Hall [213] for an in-depth explanation). Those patterns are *possible* choices posing major risks, but *not inherent* properties of Type II systems—the content of whose future novel ideas and related decisions *cannot* be prophesied a priori. In short, there is no meaningful total order of "dangerousness" according to which one can compare Type I and Type II AIs. To put it plainly: *both* the *worst risk* and the *greatest luck* for a Type II system could be a Type II system.

6. Materials and Methods

6.1. RDA Data Collection

For the collection of RDA samples utilized for illustration purposes in this paper, we undertook a simple keyword-based web search limited to articles in the period between 2018 and 2020. The main queries (with associated boolean operators) that we considered were: "artificial intelligence","AI", "autonomous", "neural network", "deepfakes", "AI" AND "bias", "AI" AND "failure", "AI" AND "security", "AI" AND "safety", "AI" AND "attack". While many terms are tailored to the type of keys represented in the taxonomy (*Ia*, *Ib*, *Ic*, *Id*) that served as basis for categorization in the RDA as introduced in Section 2, we also considered utmost *general* queries such as "artificial intelligence" in order to do justice to the eventuality that we might identify a novel entirely unexpected categorization pattern. With other words, we also foresaw the possibility of not yet encountered anomalies while analyzing the results. As briefly mentioned in Section 4.2, such a case would have been assigned to a *generic placeholder key for novel unknown patterns*. It would have called for further scrutiny and eventually for a future enlargement of the taxonomy. However, as mentioned in Section 4.2, we did not yet identify any novelty of this kind in the discussed RDA. Though, at a lower level, we discovered *atypical* instances of the pre-existing key-determined clusters. We tagged this atypicality by refering to corresponding clusters with the attribute "extra"—which was the case for the extra cluster of automated disconcertion linked to risk *Ia* and the extra cluster of automated peer pressure connected to risk *Ic*.

Self-evidently, the underlying search can be performed in a more sophisticated way in future AI observatory projects. First, a broader range of keys and combinations can be strategically devised in the light of RDA and RCRA results from a previous AI observatory iteration. Second, the efforts can be supported by web crawlers [217]. Third, this could be combined with sentiment analysis tools [218] to detect negatively polarized texts of interest for an RDA. Fourth, the creation of novel datasets for text classification [219] could be undertaken for the pre-existing keys of the taxonomy which might however remain insufficient with regard to placeholders for novel patterns. In this vein, we stress the importance of human analysts for a deep semantic understanding requiring explanatory knowledge especially when it comes to the discovery of subtle novel tendencies within superficially similar text sources. Morever, an intense examination of textual material can lead to a further disentanglement of pre-existing clusters—which could even reveal the need for a broader change of the taxonomic keys. In short, a safety-aware responsible RDA data collection pipeline is not entirely automatable and requires human-level understanding by analysts.

6.2. Interlinking RDA-Based RCRA Pre-Processing and RCRA DFs

As elucidated in Section 4.2, the preparatory procedure generating candidate RCRA clusters based on RDA instances consisted of 4 consecutive steps: (1) *taxonomization*, (2) *analytical clustering*, (3) *brute-force deliberation and threshold-based pruning* and finally (4) *assembly*. Subsequently, these RCRA clusters served as basis to generate RCRA DFs that we exemplified with short RCRA narratives instantiating these clusters as presented in Section 4.3. However, for the sake of simplicity, the exact methodological approach to *interlink* the preparatory procedure and the RCRA co-creation DF was not previously characterized. In a nutshell, we utilized a method we call *complementary cognitive co-creation* (CCC). While other methods are thinkable, we encourage considering CCC where possible for reasons described in the next paragraphs. Beforehand, we must specify that purposefully, the set of researchers involved in the preparatory procedure of the RCRA and the set of researchers performing the ensuing RCRA DFs were *disjunct*. For clarity, we refer to the former as *preparatory group* and to the latter as *executive group*. We explain how a complementary collaborative effort between these groups in the form of CCC can increase the *variety and illustrative power* of RCRA DFs.

After applying taxonomization and analytical clustering to the RDA instances, the preparatory group has been described in Section 4.2 to perform brute-force deliberation

and threshold-based pruning. While a brute-force search could appear suboptimal at first sight, we specifically considered this option in order to allow for the preparatory group to potentially be able to retrospectively *diversify* the generation of instances performed by the executive group given the RCRA clusters. This becomes possible, because whilst the preparatory group goes through every single available RDA instance, it attempts to generate an above threshold downward counterfactual that if identified can later turn out to be *utile to store*. In short, when a downward counterfactual is successfully generated for a given RDA sample, the preparatory group can not only maintain the RDA sample, but also store the generated downward counterfactual instance for later RCRA augmentation purposes. Thereby, as briefly specified, generic RCRA *clusters* were used instead of specific instances as inputs for the RCRA DFs to avoid overfitting to the idosyncrasies of unique events and possibly generate a broader variety of DF scenarios. In fact, by solely providing RCRA *clusters* to the executive group at the start of the DFs, we avoid a potentially biased negative influence by the narrow instances of the preparatory group that fulfilled a different primary function (namely the identification of above threshold patterns). To recapitulate, the preparatory procedure can be more precisely re-explained as follows: the preparatory group undergoes all 4 consecutive steps with the crucial additional detail that the brute-force deliberation and threshold-based pruning operation *also* includes *the storing of a successfully generated downward counterfactual instance* for each maintained factual RDA instance. After this pre-DF processing, the preparatory group delivers the RCRA clusters to the executive group which then engages in generating a variety of narratives instances for each obtained cluster. Post-DF, the executive group compares the generated instances with those imagined by the preparatory group pre-DF. All cases that were not yet considered by the executive group[18] but were generated by the preparatory group, are concatenated to the now augmented DFs. Duplicates are ignored.

This overall sequence of steps presents a theoretical collaborative basis for *an augmentation of co-creation DFs* to which we refer to with CCC. A further tool that may improve the efficacy of CCC is to add a functional viewpoint (i.e., related to information processing in a certain context). On closer inspection, it becomes clear why CCC can profit from a functional or/and (neuro-)cognitive [220–222] diversity of the partaking researchers. Given that in the human cognitive domain, variety is the norm [223] and heterogeneity can provide requisite variety in complex multi-causal dynamic problem domains [214] necessitating collective learning [220] and innovation [224], it makes sense to explore this potential. For instance, while the preparatory group can especially profit from individuals that excel at convergent thinking, the executive group may benefit from divergent thinkers. Pre-DF, the preparatory group needs to map from *one* factual *instance* to *one* counterfactual *instance*. In the DF, the executive group maps from *one* counterfactual *cluster* to *many* counterfactual *instances*. The former requires a horizontal integration at a low level of abstraction while the latter requires a vertical integration from a higher to a lower level of abstraction revealing the potential for *complementary* synergies[19]. A CCC-based approach combining a preparatory group comprising i.a. individuals with a cognitive profile exhibiting strengths in the former and an executive group i.a. sampled from a pool of individuals with strengths in the latter could increase efficiency, variety and illustrative power of RDA-based RCRA co-creation DFs—critical to raise safety-awareness in experts but also in the public.

---

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

[19] For instance, despite possible significant context-dependent [223] hindrances, dyadic mismatches [225] and disabilities, autistic traits are also paired with enhanced convergent thinking [226], detail-rich thinking [227] and higher verbal creativity [228] while attention deficit hyperactivity disorder traits have been linked to enhanced divergent thinking [229,230] and enhanced originality and flexibility [231]. Systematically combining these two complementary cognitive profiles under a CCC-oriented approach to RDA-based RCRA-DFs for AI observatory feedback-loops could engender benefits.

7. Conclusions

Starting with a *cybersecurity-oriented* fit-for-purpose taxonomy of ethical distinction, we introduced and exemplified a *retrospective descriptive analysis* (RDA) for future AI observatory projects. Subsequently, we elucidated how to craft a complementary *retrospective counterfactual risk analysis* (RCRA) based on downward counterfactuals from the previously extracted factual RDA samples. Motivated by recent work on risk management of hazardous events [14] and the *functional theory of counterfactual thinking* [15] from social psychology, we elaborated on why an RDA-based RCRA may be suitable for risk analyses in a complex multi-causal domain such as AI safety. Thereafter, in the light of the ethical sensitivity of AI risk instantiations, we discussed the use of harm intensity ratings for samples of an AI observatory given the perceiver-dependent, harm-based and dyadic nature of human cognitive templates in morality [25]. For illustrative purposes, we suggested a threshold-based approach focusing the RDA-based RCRA on downward counterfactuals of at least *lethal* dimensions. On the one hand, such a high threshold may engender fewer discrepancies in the moral perception being related to harm. On the other hand, it may simultaneously represent a suitable threshold reinforcing *mortality salience* (i.e., the awareness of one's mortality). From the perspective of a relevant socio-psychological theory denoted terror management theory [232,233], mortality salience—whose elicitation is also conceivable in co-creation design fictions from HCI including virtual reality settings [234]—may be able to foster safety-awareness and cautionary attitudes [234,235]. Against the backdrop of the RDA samples collected and our targeted RDA-based RCRA, we formulated the need for inherently *transdisciplinary* and *hybrid cognitive-affective* AI observatory and AI safety strategies. As guidelines for future work, we compiled a rich variety of tailored multi-level *near-term* solutions.

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

In a nutshell, we collated retrospective analyses complemented by future-oriented contradistinctions in order to: (1) apprise future AI observatory projects using *concrete examples* from practice and technically plausible above threshold downward counterfactuals, (2) thematizing possibly decisive bifurcations in future AI (safety) research and (3) pointing out the requirement of a constructive collaborative dialectical approach addressing those. As stated by Popper, *"while differing widely in the various little bits we know, in our infinite ignorance we are all equal"* [167]. Time might tell whether the assumption that *"the price of security is artificial stupidity"* or rather that *"the price of security is eternal creativity"* [13] (or none of those) turns out to practically solve long-term AI safety problems. Either way, explanatory knowledge *co-creation* can heavily influence whether we will succeed in *understanding* how to transform today's vulnerability awareness and mortality salience into the currently known or unknowable *upward* counterfactuals of our counterfactual future.

References

1. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete problems in AI safety. *arXiv* **2016**, arXiv:1606.06565.
2. Dafoe, A. AI governance: A research agenda. In *Governance of AI Program*; Future of Humanity Institute, University of Oxford: Oxford, UK, 2018.
3. Everitt, T.; Lea, G.; Hutter, M. AGI safety literature review. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 5441–5449.
4. Fjeld, J.; Achten, N.; Hilligoss, H.; Nagy, A.; Srikumar, M. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Cent. Res. Publ.* **2020**, *1*, 2–5. [CrossRef]
5. Irving, G.; Christiano, P.; Amodei, D. AI safety via debate. *arXiv* **2018**, arXiv:1805.00899.
6. Turchin, A.; Denkenberger, D.; Green, B.P. Global Solutions vs. Local Solutions for the AI Safety Problem. *Big Data Cogn. Comput.* **2019**, *3*, 16. [CrossRef]
7. The Agency for Digital Italy. Italian Observatory on Artificial Intelligence. 2020. Available online: https://ia.italia.it/en/ai-observatory/ (accessed on 25 April 2020).
8. Krausová, A. Czech Republic's AI Observatory and Forum. *Lawyer Q.* **2020**, *10*.
9. Denkfabrik. AI Observatory. Digitale Arbeitsgesellschaft. 2020. Available online: https://www.denkfabrik-bmas.de/en/projects/ai-observatory (accessed on 28 November 2020).
10. OECD.AI. OECD AI Policy Observatory. 2020. Available online: https://oecd.ai/ (accessed on 25 April 2020).
11. Yampolskiy, R.V. Predicting future AI failures from historic examples. *Foresight* **2019**, *21*, 138–152. [CrossRef]
12. McGregor, S. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *arXiv* **2020**, arXiv:2011.08512.
13. Aliman, N.M. Hybrid Cognitive-Affective Strategies for AI Safety. Ph.D. Thesis, Utrecht University, Utrecht, The Netherlands, 2020.
14. Woo, G. Downward Counterfactual Search for Extreme Events. *Front. Earth Sci.* **2019**, *7*, 340. [CrossRef]
15. Roese, N.J.; Epstude, K. The functional theory of counterfactual thinking: New evidence, new challenges, new insights. In *Advances in Experimental Social Psychology*; Elsevier: Amsterdam, The Netherlands, 2017; Volume 56, pp. 1–79.
16. Aliman, N.M.; Elands, P.; Hürst, W.; Kester, L.; Thórisson, K.R.; Werkhoven, P.; Yampolskiy, R.; Ziesche, S. Error-Correction for AI Safety. In *International Conference on Artificial General Intelligence*; Springer: Cham, Switzerland, 2020; pp. 12–22.
17. Brundage, M.; Avin, S.; Clark, J.; Toner, H.; Eckersley, P.; Garfinkel, B.; Dafoe, A.; Scharre, P.; Zeitzoff, T.; Filar, B.; et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv* **2018**, arXiv:1802.07228.
18. Pistono, F.; Yampolskiy, R.V. Unethical Research: How to Create a Malevolent Artificial Intelligence. *arXiv* **2016**, arXiv:1605.02817.
19. Aliman, N.M.; Kester, L.; Werkhoven, P.; Ziesche, S. Sustainable AI Safety? *Delphi Interdiscip. Rev. Emerg. Technol.* **2020**, *2*, 226–233.
20. Aliman, N.M.; Kester, L.; Werkhoven, P.; Yampolskiy, R. Orthogonality-based disentanglement of responsibilities for ethical intelligent systems. In *International Conference on Artificial General Intelligence*; Springer: Cham, Switzerland, 2019; pp. 22–31.
21. Cancila, D.; Gerstenmayer, J.L.; Espinoza, H.; Passerone, R. Sharpening the scythe of technological change: Socio-technical challenges of autonomous and adaptive cyber-physical systems. *Designs* **2018**, *2*, 52. [CrossRef]
22. Martin, D., Jr.; Prabhakaran, V.; Kuhlberg, J.; Smart, A.; Isaac, W.S. Extending the Machine Learning Abstraction Boundary: A Complex Systems Approach to Incorporate Societal Context. *arXiv* **2020**, arXiv:2006.09663.
23. Scott, P.J.; Yampolskiy, R.V. Classification Schemas for Artificial Intelligence Failures. *Delphi-Interdiscip. Rev. Emerg. Technol.* **2020**, *2*, 186–199. [CrossRef]
24. Gray, K.; Waytz, A.; Young, L. The moral dyad: A fundamental template unifying moral judgment. *Psychol. Inq.* **2012**, *23*, 206–215. [CrossRef]
25. Schein, C.; Gray, K. The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personal. Soc. Psychol. Rev.* **2018**, *22*, 32–70. [CrossRef]
26. Gray, K.; Schein, C.; Cameron, C.D. How to think about emotion and morality: Circles, not arrows. *Curr. Opin. Psychol.* **2017**, *17*, 41–46. [CrossRef]
27. Popper, K.R. *The Poverty of Historicism*; Routledge & Kegan Paul: Abingdon, UK, 1966.
28. Aliman, N.; Kester, L. Malicious Design in AIVR, Falsehood and Cybersecurity-oriented Immersive Defenses. In Proceedings of the 2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR), Utrecht, The Netherlands, 14–18 December 2020.

29. Harwell, D. An Artificial-Intelligence First: Voice-Mimicking Software Reportedly Used in a Major Theft. *The Washington Post*, 2019. Available online: https://www.washingtonpost.com/technology/2019/09/04/an-artificial-intelligence-first-voice-mimicking-software-reportedly-used-major-theft/ (accessed on 4 August 2020).

30. Rohrlich, J. Romance Scammer Used Deepfakes to Impersonate a Navy Admiral and Bilk Widow Out of Nearly $300,000. *Daily Beastl*, 2020. Available online: https://www.thedailybeast.com/romance-scammer-used-deepfakes-to-impersonate-a-navy-admiral-and-bilk-widow-out-of-nearly-dollar300000 (accessed on 8 November 2020).

31. Rushing, E. A Philly Lawyer Nearly Wired $9,000 to a Stranger Impersonating His Son's Voice, Showing Just How Smart Scammers are Getting. *The Philadelphia Inquirer*, 2020. Available online: https://www.inquirer.com/news/voice-scam-impersonation-fraud-bail-bond-artificial-intelligence-20200309.html (accessed on 4 August 2020).

32. Stupp, C. Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. *Wall Str. J.* **2019**. Available online: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 (accessed on 4 August 2020).

33. Gieseke, A.P. "The New Weapon of Choice": Law's Current Inability to Properly Address Deepfake Pornography. *Vanderbilt Law Rev.* **2020**, *73*, 1479–1515.

34. Ajder, H.; Patrini, G.; Cavalli, F.; Cullen, L. The State of Deepfakes: Landscape, Threats, and Impact. *Amst. Deep.* **2019**, *1*, 1–15.

35. Alba, D. Facebook Discovers Fakes That Show Evolution of Disinformation. *The New York Times*, 2019. Available online: https://www.nytimes.com/2019/12/20/business/facebook-ai-generated-profiles.html (accessed on 4 August 2020).

36. Reuters. Deepfake Used to Attack Activist Couple Shows New Disinformation Frontier. Reuters. 2020. Available online: https://gadgets.ndtv.com/internet/features/deepfake-oliver-taylor-mazen-masri-terrorist-accuse-london-university-of-birmingham-student-fake-profile-22640449 (accessed on 8 November 2020).

37. Cole, S.; Maiberg, E. Deepfake Porn Is Evolving to Give People Total Control Over Women's Bodies. *VICE*, 2019. Available online: https://www.vice.com/en/article/9keen8/deepfake-porn-is-evolving-to-give-people-total-control-over-womens-bodies (accessed on 8 November 2020).

38. Hao, K. A Deepfake Bot Is Being Used to "Undress" Underage Girls. *MIT Technol. Rev.* **2020**. Available online: https://www.technologyreview.com/2020/10/20/1010789/ai-deepfake-bot-undresses-women-and-underage-girls/ (accessed on 8 November 2020).

39. Corera, G. UK Spies will Need Artificial Intelligence—Rusi Report. *BBC*, 2020. Available online: https://www.bbc.com/news/technology-52415775 (accessed on 8 November 2020).

40. Satter, R. Experts: Spy Used AI-Generated Face to Connect With Targets. *AP News*, 2019. Available online: https://apnews.com/article/bc2f19097a4c4fffaa00de6770b8a60d (accessed on 4 August 2020).

41. Probyn, A.; Doran, M. One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority. *ABC News*, 2020. Available online: https://www.abc.net.au/news/2020-09-14/chinese-data-leak-linked-to-military-names-australians/12656668 (accessed on 4 August 2020).

42. Mozur, P. China's 'Hybrid War': Beijing's Mass Surveillance of Australia And the World for Secrets and Scandal. *The New York Times*, 2019. Available online: https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html (accessed on 4 August 2020).

43. Neekhara, P.; Hussain, S.; Jere, M.; Koushanfar, F.; McAuley, J. Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples. *arXiv* **2020**, arXiv:2002.12749.

44. Zang, J.; Sweeney, L.; Weiss, M. The Real Threat of Fake Voices in a Time of Crisis. Techcrunch. 2020. Available online: https://techcrunch.com/2020/05/16/the-real-threat-of-fake-voices-in-a-time-of-crisis/?guccounter=1 (accessed on 8 November 2020).

45. O'Donnell, L. Black Hat 2020: Open-Source AI to Spur Wave of 'Synthetic Media' Attacks. Threatpost. 2020. Available online: https://threatpost.com/black-hat-2020-open-source-ai-to-spur-wave-of-synthetic-media-attacks/158066/ (accessed on 8 November 2020).

46. Transformer, G.P., Jr.; Note, E.X.; Spellchecker, M.S.; Yampolskiy, R. When Should Co-Authorship Be Given to AI? Unpublished, PhilArchive. 2020. Available online: https://philarchive.org/archive/GPTWSCv1 (accessed on 8 November 2020).

47. Zhang, F.; Zhou, S.; Qin, Z.; Liu, J. Honeypot: A supplemented active defense system for network security. In Proceedings of the Fourth International Conference on Parallel and Distributed Computing, Applications and Technologies, Chengdu, China, 29 August 2003; pp. 231–235.

48. Nelson, S.D.; Simek, J.W. Video and Audio Deepfakes: What Lawyers Need to Know. Sensei Enterprises, Inc. 2020. Available online: https://www.masslomap.org/video-and-audio-deepfakes-what-lawyers-need-to-know-guest-post/ (accessed on 8 November 2020).

49. Chen, T.; Liu, J.; Xiang, Y.; Niu, W.; Tong, E.; Han, Z. Adversarial attack and defense in reinforcement learning-from AI security view. *Cybersecurity* **2019**, *2*, 11. [CrossRef]

50. Spocchia, G. Republican Candidate Shares Conspiracy Theory That George Floyd Murder Was Faked. Independent. 2020. Available online: https://www.independent.co.uk/news/world/americas/us-politics/george-floyd-murder-fake-conspiracy-theory-hoax-republican-gop-missouri-a9580896.html (accessed on 4 August 2020).

51. Hao, K. The Biggest Threat of Deepfakes Isn't the Deepfakes Themselves. *MIT Technol. Rev.* **2019**. Available online: https://www.technologyreview.com/2019/10/10/132667/the-biggest-threat-of-deepfakes-isnt-the-deepfakes-themselves/ (accessed on 8 November 2020).

52. Bilge, L.; Dumitraş, T. Before we knew it: An empirical study of zero-day attacks in the real world. In Proceedings of the 2012 ACM Conference on Computer and Communications Security, Raleigh, NC, USA, 16–18 October 2012; pp. 833–844.

53. Carlini, N.; Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, Dallas, TX, USA, 3 November 2017; pp. 3–14

54. Carlini, N. A Partial Break of the Honeypots Defense to Catch Adversarial Attacks. *arXiv* **2020**, arXiv:2009.10975.

55. Papernot, N.; McDaniel, P.; Sinha, A.; Wellman, M. Towards the science of security and privacy in machine learning. *arXiv* **2016**, arXiv:1611.03814.

56. Tramer, F.; Carlini, N.; Brendel, W.; Madry, A. On adaptive attacks to adversarial example defenses. *arXiv* **2020**, arXiv:2002.08347.

57. Kirat, D.; Jang, J.; Stoecklin, M. Deeplocker–Concealing Targeted Attacks with AI Locksmithing. *Blackhat USA* **2018**, *1*, 1–29.

58. Qiu, H.; Xiao, C.; Yang, L.; Yan, X.; Lee, H.; Li, B. Semanticadv: Generating adversarial examples via attribute-conditional image editing. *arXiv* **2019**, arXiv:1906.07927.

59. Carlini, N.; Farid, H. Evading Deepfake-Image Detectors with White-and Black-Box Attacks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 658–659.

60. Xu, K.; Zhang, G.; Liu, S.; Fan, Q.; Sun, M.; Chen, H.; Chen, P.Y.; Wang, Y.; Lin, X. Adversarial t-shirt! Evading person detectors in a physical world. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 665–681.

61. Wallace, E.; Feng, S.; Kandpal, N.; Gardner, M.; Singh, S. Universal Adversarial Triggers for Attacking and Analyzing NLP. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019.

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

63. Finlayson, S.G.; Bowers, J.D.; Ito, J.; Zittrain, J.L.; Beam, A.L.; Kohane, I.S. Adversarial attacks on medical machine learning. *Science* **2019**, *363*, 1287–1289. [CrossRef]

64. Han, X.; Hu, Y.; Foschini, L.; Chinitz, L.; Jankelson, L.; Ranganath, R. Deep learning models for electrocardiograms are susceptible to adversarial attack. *Nat. Med.* **2020**, *26*, 360–363. [CrossRef]

65. Zhang, X.; Wu, D.; Ding, L.; Luo, H.; Lin, C.T.; Jung, T.P.; Chavarriaga, R. Tiny noise, big mistakes: Adversarial perturbations induce errors in brain-computer interface spellers. *Natl. Sci. Rev.* **2020**, *10*, 3837.

66. Zhou, Z.; Tang, D.; Wang, X.; Han, W.; Liu, X.; Zhang, K. Invisible mask: Practical attacks on face recognition with infrared. *arXiv* **2018**, arXiv:1803.04683.

67. Cao, Y.; Xiao, C.; Cyr, B.; Zhou, Y.; Park, W.; Rampazzi, S.; Chen, Q.A.; Fu, K.; Mao, Z.M. Adversarial sensor attack on LiDAR-based perception in autonomous driving. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, London, UK, 11–15 November 2019; pp. 2267–2281.

68. Povolny, S.; Trivedi, S. Model Hacking ADAS to Pave Safer Roads for Autonomous Vehicles. McAfee. 2020. Available online: https://www.mcafee.com/blogs/other-blogs/mcafee-labs/model-hacking-adas-to-pave-safer-roads-for-autonomous-vehicles/ (accessed on 8 November 2020).

69. Chen, Y.; Yuan, X.; Zhang, J.; Zhao, Y.; Zhang, S.; Chen, K.; Wang, X. Devil's Whisper: A General Approach for Physical Adversarial Attacks against Commercial Black-box Speech Recognition Devices. In Proceedings of the 29th USENIX Security Symposium (USENIX Security 20). USENIX Association, Online, 12–14 August 2020; pp. 2667–2684.

70. Li, J.; Qu, S.; Li, X.; Szurley, J.; Kolter, J.Z.; Metze, F. Adversarial music: Real world audio adversary against wake-word detection system. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2019; pp. 11931–11941.

71. Wu, J.; Zhou, M.; Liu, S.; Liu, Y.; Zhu, C. Decision-based Universal Adversarial Attack. *arXiv* **2020**, arXiv:2009.07024.

72. Shumailov, I.; Zhao, Y.; Bates, D.; Papernot, N.; Mullins, R.; Anderson, R. Sponge Examples: Energy-Latency Attacks on Neural Networks. *arXiv* **2020**, arXiv:2006.03463.

73. Cinà, A.E.; Torcinovich, A.; Pelillo, M. A Black-box Adversarial Attack for Poisoning Clustering. *arXiv* **2020**, arXiv:2009.05474.

74. Chitpin, S. Should Popper's view of rationality be used for promoting teacher knowledge? *Educ. Philos. Theory* **2013**, *45*, 833–844. [CrossRef]

75. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **2019**, *366*, 447–453. [CrossRef] [PubMed]

76. Hill, K. Wrongfully accused by an algorithm. *The New York Times*, 24 June 2020.

77. Buolamwini, J.; Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23–24 February 2018.

78. Da Costa, C. The Women Geniuses Taking on Racial and Gender Bias in AI—And Amazon. *The Daily Beast*, 2020. Available online: https://www.thedailybeast.com/the-women-geniuses-taking-on-racial-and-gender-bias-in-artificial-intelligence-and-amazon (accessed on 23 May 2020).

79. Larrazabal, A.J.; Nieto, N.; Peterson, V.; Milone, D.H.; Ferrante, E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 12592–12594. [CrossRef]

80. Prabhu, V.U.; Birhane, A. Large image datasets: A pyrrhic win for computer vision? *arXiv* **2020**, arXiv:2006.16923.

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

82. Kempsell, R. Ofqual Pauses Study into Whether AI Could be Used to Mark Exams. *The Times*, 2020. Available online: https://www.thetimes.co.uk/article/robot-exam-marking-project-is-put-on-hold-vvrm753l3 (accessed on 10 November 2020).

83. Huchel, B. Artificial Intelligence Examines Best Ways to Keep Parolees From Recommitting Crimes. *Phys. Org.* **2020**. Available online: https://phys.org/news/2020-08-artificial-intelligence-ways-parolees-recommitting.html (accessed on 20 August 2020).

84. Cushing, T. Harrisburg University Researchers Claim Their 'Unbiased' Facial Recognition Software Can Identify Potential Criminals. techdirt. 2020. Available online: https://www.techdirt.com/articles/20200505/17090244442/harrisburg-university-researchers-claim-their-unbiased-facial-recognition-software-can-identify-potential-criminals.shtml (accessed on 2 November 2020).

85. Harrisburg University. HU Facial Recognition Software Predicts Criminality. 2020. Available online: http://archive.is/N1HVe#selection-1509.0-1509.51 (accessed on 23 May 2020).

86. Pascu, L. Biometric Software that Allegedly Predicts Criminals Based on Their Face Sparks Industry Controversy. Biometric. 2020. Available online: https://www.biometricupdate.com/202005/biometric-software-that-allegedly-predicts-criminals-based-on-their-face-sparks-industry-controversy (accessed on 23 May 2020).

87. Barrett, L.F.; Adolphs, R.; Marsella, S.; Martinez, A.M.; Pollak, S.D. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychol. Sci. Public Interest* **2019**, *20*, 1–68. [CrossRef]

88. Gendron, M.; Hoemann, K.; Crittenden, A.N.; Mangola, S.M.; Ruark, G.A.; Barrett, L.F. Emotion perception in Hadza Hunter-Gatherers. *Sci. Rep.* **2020**, *10*, 1–17. [CrossRef] [PubMed]

89. Crawford, K.; Dobbe, R.; Dryer, T.; Fried, G.; Green, B.; Kaziunas, E.; Kak, A.; Mathur, V.; McElroy, E.; Sánchez, A.N.; et al. *AI Now 2019 Report*; AI Now Institute: New York, NY, USA, 2019. Available online: https://ainowinstitute.org/AI_Now_2019_Report.pdf (accessed on 23 May 2020).

90. Lieber, C. Tech Companies Use "Persuasive Design" to Get Us Hooked. Psychologists Say It's Unethical. Vox. 2018. Available online: https://www.vox.com/2018/8/8/17664580/persuasive-technology-psychology (accessed on 8 November 2020).

91. Jakubowski, G. What's not to like? Social media as information operations force multiplier. *Jt. Force Q.* **2019**, *3*, 8–17.

92. Sawers, P. The Social Dilemma: How Digital Platforms Pose an Existential Threat to Society. VentureBeat. 2020. Available online: https://venturebeat.com/2020/09/02/the-social-dilemma-how-digital-platforms-pose-an-existential-threat-to-society/ (accessed on 2 November 2020).

93. Chikhale, S.; Gohad, V. Multidimensional Construct About The Robot Citizenship Law's In Saudi Arabia. *Int. J. Innov. Res. Adv. Stud. (IJIRAS)* **2018**, *5*, 106–108.

94. Yam, K.C.; Bigman, Y.E.; Tang, P.M.; Ilies, R.; De Cremer, D.; Soh, H.; Gray, K. Robots at work: People prefer—And forgive—Service robots with perceived feelings. *J. Appl. Psychol.* **2020**, *1*, 1–16. [CrossRef] [PubMed]

95. Orabi, M.; Mouheb, D.; Al Aghbari, Z.; Kamel, I. Detection of Bots in Social Media: A Systematic Review. *Inf. Process. Manag.* **2020**, *57*, 102250. [CrossRef]

96. Prier, J. Commanding the trend: Social media as information warfare. *Strateg. Stud. Q.* **2017**, *11*, 50–85.

97. Letter, O. Our Letter to the APA. 2018. Available online: https://screentimenetwork.org/apa (accessed on 2 November 2020).

98. Theriault, J.E.; Young, L.; Barrett, L.F. The sense of should: A biologically-based framework for modeling social pressure. *Phys. Life Rev.* **2020**, in press. [CrossRef]

99. Anderson, M.; Jiang, J. Teens' social media habits and experiences. *Pew Res. Cent.* **2018**, *28*, 1.

100. Barberá, P.; Zeitzoff, T. The new public address system: Why do world leaders adopt social media? *Int. Stud. Q.* **2018**, *62*, 121–130. [CrossRef]

101. Franchina, V.; Coco, G.L. The influence of social media use on body image concerns. *Int. J. Psychoanal. Educ.* **2018**, *10*, 5–14.

102. Halfmann, A.; Rieger, D. Permanently on call: The effects of social pressure on smartphone users' self-control, need satisfaction, and well-being. *J. Comput. Mediat. Commun.* **2019**, *24*, 165–181. [CrossRef]

103. Stieger, S.; Lewetz, D. A week without using social media: Results from an ecological momentary intervention study using smartphones. *Cyberpsychol. Behav. Soc. Netw.* **2018**, *21*, 618–624. [CrossRef] [PubMed]

104. Ferrara, E.; Yang, Z. Measuring emotional contagion in social media. *PLoS ONE* **2015**, *10*, e0142390. [CrossRef] [PubMed]

105. Luxton, D.D.; June, J.D.; Fairall, J.M. Social media and suicide: A public health perspective. *Am. J. Public Health* **2012**, *102*, S195–S200. [CrossRef] [PubMed]

106. Lane, L. NIST finds flaws in facial checks on people with Covid masks. *Biom. Technol. Today* **2020**, *8*, 2.

107. Mundial, I.Q.; Hassan, M.S.U.; Tiwana, M.I.; Qureshi, W.S.; Alanazi, E. Towards Facial Recognition Problem in COVID-19 Pandemic. In Proceedings of the 2020 4rd International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM), Medan, Indonesia, 3–4 September 2020; pp. 210–214.

108. Ngan, M.L.; Grother, P.J.; Hanaoka, K.K. Ongoing Face Recognition Vendor Test (FRVT) Part 6A: Face recognition accuracy with masks using pre-COVID-19 algorithms. *Natl. Inst. Stand. Technol.* **2020**, *1*, 1.

109. Krishna, K.; Tomar, G.S.; Parikh, A.P.; Papernot, N.; Iyyer, M. Thieves on Sesame Street! Model Extraction of BERT-based APIs. *arXiv* **2019**, arXiv:1910.12366.

110. Taylor, J. Facebook Incorrectly Removes Picture of Aboriginal Men in Chains Because of 'Nudity'. *The Guardian*, 2020. Available online: https://www.theguardian.com/technology/2020/jun/13/facebook-incorrectly-removes-picture-of-aboriginal-men-in-chains-because-of-nudity (accessed on 2 November 2020).

111. DeCamp, M.; Lindvall, C. Latent bias and the implementation of artificial intelligence in medicine. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 2020–2023. [CrossRef]

112. Kaushal, A.; Altman, R.; Langlotz, C. Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. *JAMA* **2020**, *324*, 1212–1213. [CrossRef]

113. Epstude, K.; Roese, N.J. The functional theory of counterfactual thinking. *Personal. Soc. Psychol. Rev.* **2008**, *12*, 168–192. [CrossRef]

114. Weidman, G. *Penetration Testing: A Hands-On Introduction to Hacking*; No Starch Press: San Francisco, CA, USA, 2014.

115. Rajendran, J.; Jyothi, V.; Karri, R. Blue team red team approach to hardware trust assessment. In Proceedings of the 2011 IEEE 29th International Conference on Computer Design (ICCD), Amherst, MA, USA, 9–12 October 2011; pp. 285–288.

116. Rege, A. Incorporating the human element in anticipatory and dynamic cyber defense. In Proceedings of the 2016 IEEE International Conference on Cybercrime and Computer Forensic (ICCCF), Vancouver, BC, Canada, 12–14 June 2016; pp. 1–7.

117. Ahmadpour, N.; Pedell, S.; Mayasari, A.; Beh, J. Co-creating and assessing future wellbeing technology using design fiction. *She Ji J. Des. Econ. Innov.* **2019**, *5*, 209–230. [CrossRef]

118. Pillai, A.G.; Ahmadpour, N.; Yoo, S.; Kocaballi, A.B.; Pedell, S.; Sermuga Pandian, V.P.; Suleri, S. Communicate, Critique and Co-create (CCC) Future Technologies through Design Fictions in VR Environment. In Proceedings of the Companion Publication of the 2020 ACM Designing Interactive Systems Conference, Eindhoven, The Netherlands, 6–20 July 2020; pp. 413–416.

119. Rapp, A. Design fictions for learning: A method for supporting students in reflecting on technology in Human-Computer Interaction courses. *Comput. Educ.* **2020**, *145*, 103725. [CrossRef]

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

121. Carlini, N.; Athalye, A.; Papernot, N.; Brendel, W.; Rauber, J.; Tsipras, D.; Goodfellow, I.; Madry, A.; Kurakin, A. On evaluating adversarial robustness. *arXiv* **2019**, arXiv:1902.06705.

122. John, A.; Glendenning, A.C.; Marchant, A.; Montgomery, P.; Stewart, A.; Wood, S.; Lloyd, K.; Hawton, K. Self-harm, suicidal behaviours, and cyberbullying in children and young people: Systematic review. *J. Med. Internet Res.* **2018**, *20*, e129. [CrossRef]

123. Crothers, B. FBI Warns on Teenage Sextortion as New Twists on Sex-Related Scams Emerge. *Fox News*, 2020. Available online: https://www.foxnews.com/tech/fbi-warns-teenage-sextortion-new-twists-sex-scams-emerge (accessed on 2 November 2020).

124. Nilsson, M.G.; Pepelasi, K.T.; Ioannou, M.; Lester, D. Understanding the link between Sextortion and Suicide. *Int. J. Cyber Criminol.* **2019**, *13*, 55–69.

125. Haag, M.; Salam, M. Gunman in 'Pizzagate' Shooting Is Sentenced to 4 Years in Prison. *The New York Times*, 2017. Available online: https://www.nytimes.com/2017/06/22/us/pizzagate-attack-sentence.html (accessed on 2 November 2017).

126. Bessi, A.; Ferrara, E. Social bots distort the 2016 US Presidential election online discussion. *First Monday* **2016**, *21*, 1–14.

127. Assenmacher, D.; Clever, L.; Frischlich, L.; Quandt, T.; Trautmann, H.; Grimme, C. Demystifying Social Bots: On the Intelligence of Automated Social Media Actors. *Soc. Media Soc.* **2020**, *6*, 2056305120939264. [CrossRef]

128. Boneh, D.; Grotto, A.J.; McDaniel, P.; Papernot, N. How relevant is the Turing test in the age of sophisbots? *IEEE Secur. Priv.* **2019**, *17*, 64–71. [CrossRef]

129. Yang, K.C.; Varol, O.; Davis, C.A.; Ferrara, E.; Flammini, A.; Menczer, F. Arming the public with artificial intelligence to counter social bots. *Hum. Behav. Emerg. Technol.* **2019**, *1*, 48–61. [CrossRef]

130. Shao, C.; Ciampaglia, G.L.; Varol, O.; Yang, K.C.; Flammini, A.; Menczer, F. The spread of low-credibility content by social bots. *Nat. Commun.* **2018**, *9*, 1–9. [CrossRef] [PubMed]

131. Yan, H.Y.; Yang, K.C.; Menczer, F.; Shanahan, J. Asymmetrical perceptions of partisan political bots. *New Media Soc.* **2020**. [CrossRef]

132. Farokhmanesh, M. Is It Legal to Swap Someone's Face into Porn without Consent? *Verge. January* **2018**, *30*, 1.

133. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and improving the image quality of StyleGAN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 8110–8119.

134. Duan, R.; Ma, X.; Wang, Y.; Bailey, J.; Qin, A.K.; Yang, Y. Adversarial Camouflage: Hiding Physical-World Attacks with Natural Styles. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 1000–1008.

135. Kong, Z.; Guo, J.; Li, A.; Liu, C. PhysGAN: Generating Physical-World-Resilient Adversarial Examples for Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 14254–14263.

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

137. Wang, Y.; Lv, H.; Kuang, X.; Zhao, G.; Tan, Y.A.; Zhang, Q.; Hu, J. Towards a Physical-World Adversarial Patch for Blinding Object Detection Models. *Inf. Sci.* **2020**, in press. [CrossRef]

138. Rahman, A.; Hossain, M.S.; Alrajeh, N.A.; Alsolami, F. Adversarial examples–security threats to COVID-19 deep learning systems in medical IoT devices. *IEEE Internet Things J.* **2020**. [CrossRef]

139. Ciosek, I. Aggravating Uncertaub–Russian Information Warfare in the West. *Tor. Int. Stud.* **2020**, *1*, 57–72. [CrossRef]
140. Colleoni, E.; Rozza, A.; Arvidsson, A. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *J. Commun.* **2014**, *64*, 317–332. [CrossRef]
141. Kocabey, E.; Ofli, F.; Marin, J.; Torralba, A.; Weber, I. Using computer vision to study the effects of BMI on online popularity and weight-based homophily. In *International Conference on Social Informatics*; Springer: Cham, Switzerland, 2018; pp. 129–138.
142. Hanusch, F.; Nölleke, D. Journalistic homophily on social media: Exploring journalists' interactions with each other on Twitter. *Digit. J.* **2019**, *7*, 22–44. [CrossRef]
143. Lathiya, S.; Dhobi, J.; Zubiaga, A.; Liakata, M.; Procter, R. Birds of a feather check together: Leveraging homophily for sequential rumour detection. *Online Soc. Netw. Media* **2020**, *19*, 100097. [CrossRef]
144. Leonhardt, J.M.; Pezzuti, T.; Namkoong, J.E. We're not so different: Collectivism increases perceived homophily, trust, and seeking user-generated product information. *J. Bus. Res.* **2020**, *112*, 160–169. [CrossRef]
145. Saleem, A.; Ellahi, A. Influence of electronic word of mouth on purchase intention of fashion products in social networking websites. *Pak. J. Commer. Soc. Sci. (PJCSS)* **2017**, *11*, 597–622.
146. Ismagilova, E.; Slade, E.; Rana, N.P.; Dwivedi, Y.K. The effect of characteristics of source credibility on consumer behaviour: A meta-analysis. *J. Retail. Consum. Serv.* **2020**, *53*, 1–9. [CrossRef]
147. Kim, S.; Kandampully, J.; Bilgihan, A. The influence of eWOM communications: An application of online social network framework. *Comput. Hum. Behav.* **2018**, *80*, 243–254. [CrossRef]
148. Ladhari, R.; Massa, E.; Skandrani, H. YouTube vloggers' popularity and influence: The roles of homophily, emotional attachment, and expertise. *J. Retail. Consum. Serv.* **2020**, *54*, 102027. [CrossRef]
149. Xu, S.; Zhou, A. Hashtag homophily in twitter network: Examining a controversial cause-related marketing campaign. *Comput. Hum. Behav.* **2020**, *102*, 87–96. [CrossRef]
150. Zhou, Z.; Xu, K.; Zhao, J. Homophily of music listening in online social networks of China. *Soc. Netw.* **2018**, *55*, 160–169. [CrossRef]
151. Vonk, R. Effects of stereotypes on attitude inference: Outgroups are black and white, ingroups are shaded. *Br. J. Soc. Psychol.* **2002**, *41*, 157–167. [CrossRef]
152. Bakshy, E.; Messing, S.; Adamic, L.A. Exposure to ideologically diverse news and opinion on Facebook. *Science* **2015**, *348*, 1130–1132. [CrossRef]
153. Lamb, A. After Covid, AI Will Pivot. Towards Data Sciece. 2020. Available online: https://towardsdatascience.com/after-covid-ai-will-pivot-dbe9dd06327 (accessed on 12 November 2020).
154. Smith, G.; Rustagi, I. The Problem With COVID-19 Artificial Intelligence Solutions and How to Fix Them. Standford Social Innovation Review. 2020. Available online: https://ssir.org/articles/entry/the_problem_with_covid_19_artificial_intelligence_solutions_and_how_to_fix_them (accessed on 12 November 2020).
155. Yampolskiy, R.V. Mimicry attack on strategy-based behavioral biometric. In Proceedings of the Fifth International Conference on Information Technology: New Generations (ITNG 2008), Las Vegas, NV, USA, 7–9 April 2008; pp. 916–921.
156. Yampolskiy, R.V.; Govindaraju, V. Taxonomy of behavioural biometrics. In *Behavioral Biometrics for Human Identification: Intelligent Applications*; IGI Global: Hershey, PA, USA, 2010; pp. 1–43.
157. Yampolskiy, R.V. Analyzing user password selection behavior for reduction of password space. In Proceedings of the 40th Annual 2006 International Carnahan Conference on Security Technology, Lexington, KY, USA, 16–19 October 2006; pp. 109–115.
158. Whyte, C. Deepfake news: AI-enabled disinformation as a multi-level public policy challenge. *J. Cyber Policy* **2020**, *5*, 199–217. [CrossRef]
159. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2014; pp. 2672–2680.
160. Tucciarelli, R.; Vehar, N.; Tsakiris, M. On the realness of people who do not exist: The social processing of artificial faces. *PsyArXiv* **2020**.
161. Young, L. Calibration Camouflage: Hyphen-Labs and Adam Harvey: HyperFace. *Archit. Des.* **2019**, *89*, 28–31. [CrossRef]
162. Baggili, I.; Behzadan, V. Founding The Domain of AI Forensics. *arXiv* **2019**, arXiv:1912.06497.
163. Schneider, J.; Breitinger, F. AI Forensics: Did the Artificial Intelligence System Do It? Why? *arXiv* **2020**, arXiv:2005.13635.
164. Rosenberg, A.A.; Halpern, M.; Shulman, S.; Wexler, C.; Phartiyal, P. Reinvigorating the role of science in democracy. *PLoS Biol.* **2013**, *11*, e1001553. [CrossRef] [PubMed]
165. MIT Open Learning. Tackling the Misinformation Epidemic with "In Event of Moon Disaster". *MIT News*, 2020. Available online: https://news.mit.edu/2020/mit-tackles-misinformation-in-event-of-moon-disaster-0720 (accessed on 11 October 2020).
166. Fallis, D. The Epistemic Threat of Deepfakes. *Philos. Technol.* **2020**, 1–21. [CrossRef] [PubMed]
167. Popper, K. *Conjectures and Refutations: The Growth of Scientific Knowledge*; Routledge: Abingdon, UK, 2014.
168. Deutsch, D. *The Beginning of Infinity: Explanations that Transform the World*; Penguin: London, UK, 2011.
169. Baudrillard, J. *Simulacra and Simulation*; University of Michigan Press: Ann Arbor, MI, USA, 1994.
170. Hopf, H.; Krief, A.; Mehta, G.; Matlin, S.A. Fake science and the knowledge crisis: Ignorance can be fatal. *R. Soc. Open Sci.* **2019**, *6*, 190161. [CrossRef]
171. D'Amour, A.; Heller, K.; Moldovan, D.; Adlam, B.; Alipanahi, B.; Beutel, A.; Chen, C.; Deaton, J.; Eisenstein, J.; Hoffman, M.D.; et al. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *arXiv* **2020**, arXiv:2011.03395.

172. Oughton, E.J.; Ralph, D.; Pant, R.; Leverett, E.; Copic, J.; Thacker, S.; Dada, R.; Ruffle, S.; Tuveson, M.; Hall, J.W. Stochastic Counterfactual Risk Analysis for the Vulnerability Assessment of Cyber-Physical Attacks on Electricity Distribution Infrastructure Networks. *Risk Anal.* **2019**, *39*, 2012–2031. [CrossRef]

173. Almasoud, A.S.; Hussain, F.K.; Hussain, O.K. Smart contracts for blockchain-based reputation systems: A systematic literature review. *J. Netw. Comput. Appl.* **2020**, *170*, 102814. [CrossRef]

174. Cresci, S. A decade of social bot detection. *Commun. ACM* **2020**, *63*, 72–83. [CrossRef]

175. Cresci, S.; Di Pietro, R.; Petrocchi, M.; Spognardi, A.; Tesconi, M. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, 3–7 April 2017; pp. 963–972.

176. Barrett, L.F. The theory of constructed emotion: An active inference account of interoception and categorization. *Soc. Cogn. Affect. Neurosci.* **2017**, *12*, 1–23. [CrossRef] [PubMed]

177. Aliman, N.M. Self-Shielding Worlds. 2020. Available online: https://nadishamarie.jimdo.com/clipboard/ (accessed 23 November 2020).

178. Turing, I.B.A. Computing machinery and intelligence-AM Turing. *Mind* **1950**, *59*, 433. [CrossRef]

179. Pantserev, K.A. The Malicious Use of AI-Based Deepfake Technology as the New Threat to Psychological Security and Political Stability. In *Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 37–55.

180. Öhman, C. Introducing the pervert's dilemma: A contribution to the critique of Deepfake Pornography. *Ethics Inf. Technol.* **2019**, 1–8. [CrossRef]

181. Macaulay, T. New AR App will Let You Model a Virtual Companion on Anyone You Want. 2020. Available online: https://thenextweb.com/neural/2020/06/01/new-ar-app-will-let-you-model-a-virtual-companion-on-anyone-you-want/ (accessed on 4 August 2020).

182. Kumar, R.S.S.; Nyström, M.; Lambert, J.; Marshall, A.; Goertzel, M.; Comissoneru, A.; Swann, M.; Xia, S. Adversarial Machine Learning–Industry Perspectives. *arXiv* **2020**, arXiv:2002.05646.

183. Barrett, L.F.; Simmons, W.K. Interoceptive predictions in the brain. *Nat. Rev. Neurosci.* **2015**, *16*, 419–429. [CrossRef]

184. Kleckner, I.R.; Zhang, J.; Touroutoglou, A.; Chanes, L.; Xia, C.; Simmons, W.K.; Quigley, K.S.; Dickerson, B.C.; Barrett, L.F. Evidence for a large-scale brain system supporting allostasis and interoception in humans. *Nat. Hum. Behav.* **2017**, *1*, 1–14. [CrossRef]

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

186. Dignum, V. AI is multidisciplinary. *AI Matters* **2020**, *5*, 18–21. [CrossRef]

187. Floridi, L. Establishing the rules for building trustworthy AI. *Nat. Mach. Intell.* **2019**, *1*, 261–262. [CrossRef]

188. Hagendorff, T. The ethics of Ai ethics: An evaluation of guidelines. *Minds Mach.* **2020**, 1–22. [CrossRef]

189. Mittelstadt, B. AI Ethics–Too principled to fail. *arXiv* **2019**, arXiv:1906.06668.

190. Whittlestone, J.; Nyrup, R.; Alexandrova, A.; Cave, S. The role and limits of principles in AI ethics: Towards a focus on tensions. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, 27–28 January 2019; pp. 195–200.

191. Jobin, A.; Ienca, M.; Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399. [CrossRef]

192. Gu, B.; Konana, P.; Raghunathan, R.; Chen, H.M. Research note—The allure of homophily in social media: Evidence from investor responses on virtual communities. *Inf. Syst. Res.* **2014**, *25*, 604–617. [CrossRef]

193. Yoo, J. Ideological Homophily and Echo Chamber Effect in Internet and Social Media. *Stud. Int. J. Res.* **2007**, *4*, 1–7.

194. Tsao, J.; Ting, C.; Johnson, C. Creative outcome as implausible utility. *Rev. Gen. Psychol.* **2019**, *23*, 279–292. [CrossRef]

195. Yampolskiy, R.V.; Ashby, L.; Hassan, L. Wisdom of Artificial Crowds—A Metaheuristic Algorithm for Optimization. *J. Intell. Learn. Syst. Appl.* **2012**, *4*, 98–107. [CrossRef]

196. Yampolskiy, R. Usable Guidelines Aim to Make AI Safer. All, EIT 2020: The Intelligent Revolution. 2020. Available online: https://www.mouser.com/blog/usable-guidelines-aim-to-make-ai-safer (accessed on 13 November 2020).

197. Trazzi, M.; Yampolskiy, R.V. Building safer AGI by introducing artificial stupidity. *arXiv* **2018**, arXiv:1808.03644.

198. Trazzi, M.; Yampolskiy, R.V. Artificial Stupidity: Data We Need to Make Machines Our Equals. *Patterns* **2020**, *1*, 100021. [CrossRef]

199. Aliman, N.M.; Kester, L. Artificial creativity augmentation. In *International Conference on Artificial General Intelligence*; Springer: Cham, Switzerland, 2020; pp. 23–33.

200. Leviathan, Y.; Matias, Y. Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone. 2018. Available online: https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html (accessed on 4 August 2020).

201. Yampolskiy, R.V. On Controllability of AI. *arXiv* **2020**, arXiv:2008.04071.

202. Yampolskiy, R.V. Unpredictability of AI. *arXiv* **2019**, arXiv:1905.13053.

203. Boström, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2014.

204. Baum, S.; Barrett, A.; Yampolskiy, R.V. Modeling and interpreting expert disagreement about artificial superintelligence. *Informatica* **2017**, *41*, 419–428.

205. Friston, K. Am I self-conscious? (Or does self-organization entail self-consciousness?). *Front. Psychol.* **2018**, *9*, 579. [CrossRef] [PubMed]

206. Bruineberg, J.; Kiverstein, J.; Rietveld, E. The anticipating brain is not a scientist: The free-energy principle from an ecological-enactive perspective. *Synthese* **2018**, *195*, 2417–2444. [CrossRef] [PubMed]

207. Rudrauf, D.; Bennequin, D.; Granic, I.; Landini, G.; Friston, K.; Williford, K. A mathematical model of embodied consciousness. *J. Theor. Biol.* **2017**, *428*, 106–131. [CrossRef] [PubMed]

208. Williford, K.; Bennequin, D.; Friston, K.; Rudrauf, D. The projective consciousness model and phenomenal selfhood. *Front. Psychol.* **2018**, *9*, 2571. [CrossRef]

209. Deutsch, D. Constructor theory. *Synthese* **2013**, *190*, 4331–4359. [CrossRef]

210. Deutsch, D.; Marletto, C. Constructor theory of information. *Proc. R. Soc. Math. Phys. Eng. Sci.* **2015**, *471*, 20140540. [CrossRef]

211. Dietrich, A. *How Creativity Happens in the Brain*; Springer: Berlin/Heidelberg, Germany, 2015.

212. Brockman, J. *Possible Minds: Twenty-Five Ways of Looking at AI. Beyond Reward and Punishment. David Deutsch.*; Penguin Books: London, UK, 2020.

213. Hall, B. Superintelligence. Part 6: Neologisms and Choices. 2020. Available online: http://www.bretthall.org/superintelligence-6.html (accessed on 4 January 2021).

214. Reynolds, A.; Lewis, D. Teams solve problems faster when they're more cognitively diverse. *Harv. Bus. Rev.* **2017**, *30*, 1–8.

215. Friston, K.J.; Lin, M.; Frith, C.D.; Pezzulo, G.; Hobson, J.A.; Ondobaka, S. Active inference, curiosity and insight. *Neural Comput.* **2017**, *29*, 2633–2683. [CrossRef]

216. Sajid, N.; Ball, P.J.; Friston, K.J. Active inference: Demystified and compared. *arXiv* **2019**, arXiv:1909.10863.

217. Hernandez, J.; Marin-Castro, H.M.; Morales-Sandoval, M. A Semantic Focused Web Crawler Based on a Knowledge Representation Schema. *Appl. Sci.* **2020**, *10*, 3837. [CrossRef]

218. Singh, N.K.; Tomar, D.S.; Sangaiah, A.K. Sentiment analysis: A review and comparative analysis over social media. *J. Ambient. Intell. Humaniz. Comput.* **2020**, *11*, 97–117. [CrossRef]

219. Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text classification algorithms: A survey. *Information* **2019**, *10*, 150. [CrossRef]

220. Aggarwal, I.; Woolley, A.W.; Chabris, C.F.; Malone, T.W. The impact of cognitive style diversity on implicit learning in teams. *Front. Psychol.* **2019**, *10*, 112. [CrossRef] [PubMed]

221. den Houting, J. Neurodiversity: An insider's perspective. *Autism* **2019**, *23*, 271–273. [CrossRef]

222. Blume, H. Neurodiversity: On the neurological underpinnings of geekdom. *Atlantic* **1998**, *30*, 1.

223. Chapman, R. Neurodiversity, disability, wellbeing. In *Neurodiversity Studies: A New Critical Paradigm*; Routledge: Abingdon, UK, 2020.

224. Chen, X.; Liu, J.; Zhang, H.; Kwan, H.K. Cognitive diversity and innovative work behaviour: The mediating roles of task reflexivity and relationship conflict and the moderating role of perceived support. *J. Occup. Organ. Psychol.* **2019**, *92*, 671–694. [CrossRef]

225. Bolis, D.; Balsters, J.; Wenderoth, N.; Becchio, C.; Schilbach, L. Beyond autism: Introducing the dialectical misattunement hypothesis and a bayesian account of intersubjectivity. *Psychopathology* **2017**, *50*, 355–372. [CrossRef]

226. Abu-Akel, A.; Webb, M.E.; de Montpellier, E.; Von Bentivegni, S.; Luechinger, L.; Ishii, A.; Mohr, C. Autistic and positive schizotypal traits respectively predict better convergent and divergent thinking performance. *Think. Ski. Creat.* **2020**, *36*, 100656. [CrossRef]

227. Paola, P.; Laura, G.; Giusy, M.; Michela, C. Autism, autistic traits and creativity: A systematic review and meta-analysis. *Cogn. Process.* **2020**, *1*, 1–36. [CrossRef]

228. Kasirer, A.; Adi-Japha, E.; Mashal, N. Verbal and Figural Creativity in Children With Autism Spectrum Disorder and Typical Development. *Front. Psychol.* **2020**, *11*, 2968. [CrossRef] [PubMed]

229. Hoogman, M.; Stolte, M.; Baas, M.; Kroesbergen, E. Creativity and ADHD: A review of behavioral studies, the effect of psychostimulants and neural underpinnings. *Neurosci. Biobehav. Rev.* **2020**, *119*, 66–85 . [CrossRef] [PubMed]

230. White, H.A. Thinking "Outside the Box": Unconstrained Creative Generation in Adults with Attention Deficit Hyperactivity Disorder. *J. Creat. Behav.* **2020**, *54*, 472–483. [CrossRef]

231. White, H.A.; Shah, P. Scope of semantic activation and innovative thinking in college students with ADHD. *Creat. Res. J.* **2016**, *28*, 275–282. [CrossRef]

232. Greenberg, J.; Arndt, J. Terror management theory. In *Handbook of Theories of Social Psychology*; SAGE Publications Inc.: Los Angeles, CA, USA, 2011; Volume 1, pp. 398–415.

233. Solomon, S.; Greenberg, J.; Pyszczynski, T. *The Worm at the Core: On the Role of Death in Life*; Random House Inc.: New York, NY, USA, 2015.

234. Chittaro, L.; Sioni, R.; Crescentini, C.; Fabbro, F. Mortality salience in virtual reality experiences and its effects on users' attitudes towards risk. *Int. J. Hum. Comput. Stud.* **2017**, *101*, 10–22. [CrossRef]

235. Shehryar, O.; Hunt, D.M. A terror management perspective on the persuasiveness of fear appeals. *J. Consum. Psychol.* **2005**, *15*, 275–287. [CrossRef]