IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 June 2022, accepted 11 July 2022, date of publication 18 July 2022, date of current version 27 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3191790

RESEARCH ARTICLE

Artificial Intelligence Crime: An Overview of Malicious Use and Abuse of AI

TAÍS FERNANDA BLAUTH, OSKAR JOSEF GSTREIN, AND ANDREJ ZWITTER
Department of Governance and Innovation, University of Groningen/Campus Fryslân, 8911 CE Leeuwarden, The Netherlands

Corresponding author: Taís Fernanda Blauth (t.f.blauth@rug.nl)

ABSTRACT The capabilities of Artificial Intelligence (AI) evolve rapidly and affect almost all sectors of society. AI has been increasingly integrated into criminal and harmful activities, expanding existing vulnerabilities, and introducing new threats. This article reviews the relevant literature, reports, and representative incidents which allows to construct a typology of the malicious use and abuse of systems with AI capabilities. The main objective is to clarify the types of activities and corresponding risks. Our starting point is to identify the vulnerabilities of AI models and outline how malicious actors can abuse them. Subsequently, we explore AI-enabled and AI-enhanced attacks. While we present a comprehensive overview, we do not aim for a conclusive and exhaustive classification. Rather, we provide an overview of the risks of enhanced AI application, that contributes to the growing body of knowledge on the issue. Specifically, we suggest four types of malicious abuse of AI (integrity attacks, unintended AI outcomes, algorithmic trading, membership inference attacks) and four types of malicious use of AI (social engineering, misinformation/fake news, hacking, autonomous weapon systems). Mapping these threats enables advanced reflection of governance strategies, policies, and activities that can be developed or improved to minimize risks and avoid harmful consequences. Enhanced collaboration among governments, industries, and civil society actors is vital to increase preparedness and resilience against malicious use and abuse of AI.

INDEX TERMS Artificial intelligence, artificial intelligence typology, computer crime, malicious artificial intelligence, security, social implications of technology.

I. INTRODUCTION

The impact of systems using Artificial Intelligence (AI) is at the center of numerous academic studies [1]–[3], political debates [4], and reports of civil society organizations [5]. The development of AI has become the subject of praise due to unprecedented technological capabilities, such as enhanced possibilities for automated image recognition (e.g., detection of cancer in the field of medicine [6], [7]). However, it has also been criticized - even feared - due to aspects such as the uncertain consequences of automation for the labor market (e.g., concerns of mass unemployment [8, pp. 26–27]). This duality of positive *vs* negative aspects of the technology can also be identified in the context of cybersecurity and cybercrime. Governments use AI to enhance their capabilities, whereas the same technology can be used for attacks against them [9].

While the recent surge in AI development has been fueled by the private sector and applications in customer-oriented applications, sectors such as defense might use similar capabilities in their operations [10]. At the same time, it is increasingly difficult to distinguish between the actions of state and non-state actors. This has recently been demonstrated by a wave of ransomware attacks targeting public infrastructure in many countries, such as the Colonial Pipeline in the United States in May 2021 [11, pp. 127–128]. Additionally, programs and applications developed for non-malicious purposes can also be implemented or modified for malicious intent and potentially cause harm.

The dual-use aspect of technology is not an entirely new problem when it comes to cybercrime[1] or (cyber-)security. Nevertheless, how AI can be leveraged for malicious use

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Wang.

[1] For the purposes of this paper, we use ''cybercrime'' in a broad sense. It includes criminal activities against data, fraud, forgery, among others [12], which can take place across borders and affect victims in different locations [13].

and abuse constitutes novel vulnerabilities. Permanent assessment of the threat landscape is crucial to create and adapt governance mechanisms, develop proactive measures, and enhance (cyber-)resilience.

To build on previous work [14]–[16] and expand the understanding of how AI broadens the potential for malicious activities online, this article evaluates the main categories of use and abuse of AI in a criminal context. We provide several salient examples that allow us to illustrate the challenges at hand. Based on these examples, we present a typology that catalogs the main harmful AI-based activities. Developing knowledge and understanding about the potential malicious use and abuse of AI enables cybersecurity organizations and governmental agencies to anticipate such incidents and increase their preparedness against attacks. Furthermore, a typology is greatly useful in structuring research efforts and identifying gaps in knowledge in areas where more research is warranted.

II. AIM OF THE STUDY AND CONTRIBUTIONS

To establish adequate security measures against AI-related attacks, it is necessary to comprehend the different types of malicious use and abuse of AI and map it, including the corresponding risks. However, there is a general lack of comprehensive and interdisciplinary assessment of the types of AI-enabled and AI-dependent cyberattacks, which might negatively affect the development of measures against them. Consequently, data security, personal safety, and political stability are at stake. This study attempts to classify different types of malicious AI to expand the body of knowledge on the subject in a more holistic manner.

Specifically, this research aims to propose a typology of the malicious use and abuse of AI based on empirical evidence and contemporary discourse, analyzing how AI systems are used to compromise confidentiality, integrity, and data availability. The technique of classification of similar subjects into groups has been established for more than 2000 years [17], and such a study can be the starting point for the development of granular and in-depth analysis. Thus, our objectives are limited to identifying essential elements of the malicious use and abuse of AI, and to collect evidence of their use in practice. The compiled data enable further analysis of the possible ways in which AI systems can be exploited for criminal activities. This research does not focus on developing a theory of malicious use and abuse of AI.

With the typology presented in this paper, we hope to make the following contributions:

a. Add to the emerging body of knowledge that maps types of malicious use and abuse of AI systems. To understand the main concepts, threat scenarios, and possibilities is necessary to develop much-needed preventive measures and proactive responses to such attacks.
b. Help in establishing a shared language among and across different disciplines, especially between STEM disciplines and legal practitioners, as well as policymakers. Interdisciplinary research on the topic can reduce confusion caused by excessively technical or monodisciplinary language and aid in bridging existing gaps.
c. Propose mitigation strategies, as well as demonstrating that a collective effort among government, academia, and industry is needed.

III. METHODOLOGY

This study refers to the categorization system of a "typology" rather than a taxonomy. The main difference between typologies and taxonomies involves the research methods used in their development: "typologies classify subjects by forcing deductive assignment into *a priori* predefined groups, while taxonomies determine membership into *a posteriori* categories that emerge from empirical analysis inductively" [17, p. 12]. Therefore, even though the terms taxonomy and typology have been used interchangeably in the literature at times [18], [19], this article refers to the classification scheme of malicious use and abuse of AI as a typology.

The methodology is based on an analysis of the available literature on cybercrime and the potential malicious use and abuse of AI systems. A literature review informs this study and findings using the following databases: IEEE Xplore, Science Direct, Wiley Online Library, and Google Scholar. We used keywords, titles, and screened abstracts. The search terms included are (Artificial Intelligence OR AI OR Machine Learning OR ML) AND (malicious OR crime OR harmful OR cyberattack). Additionally, we examined lists of references obtained from reviewed papers and reports, as well as news sources describing past AI incidents. We only reviewed papers/reports/web pages available in English and Portuguese. After analyzing these sources, we were able to identify the different types of malicious use and abuse of AI systems.

IV. DEFINITIONS

There is still no universal definition of AI. Recently, the European Commission has attempted a legal definition with the presentation of Article 3 paragraph 1 of the 2021 proposal for a European Union (EU) Regulation [20], also known as EU AI Act. The EU AI Draft Act states that the term AI system "means software that is developed with one or more [...] techniques [...] and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with." Annex I of the EU proposal elaborates on the techniques by referring to (a) Machine learning approaches,[2] (b) Logic- and knowledge-based approaches and (c) Statistical approaches, Bayesian estimation, search and optimization methods [21]. This proposed definition is subject to ongoing scrutiny by policymakers and at the time of writing it is unclear whether and in which form it

[2] In many cases, terms such as AI and Machine Learning (ML) are used interchangeably. In this paper, we use "AI" when discussing broadly and "ML" when referring to Machine Learning as an AI technique.

might continue to exist. The use of AI can enable already existing forms of crime ('cyber-enabled crime'), or establish new forms of crime ('cyber-dependent crime') [22], [23]. AI potentially enables attacks that are larger in scale and reach than previously possible with other technologies.

This article uses the term "malicious use and abuse" of AI. [15] proposed the term "AI-Crime" to describe the situation in which AI technologies are re-oriented to facilitate criminal activity. AI-Crime focuses on behavior already defined as criminal within the given legislation. We submit that this term is too limited to build a typology due to the broad scope of our analysis, which is not limited to acts that constitute a crime in each State. For instance, the creation and spread of misinformation/fake news might be harmful, but not necessarily a crime, according to certain domestic legislation. Other authors have proposed the terms "harmful AI" [24], [25] and "malevolent AI" [26]. These terms were used in a context in which the AI program/application itself caused harm. Since our analysis also includes the use of AI by individuals and organizations with the intent of causing harm, they are not suitable either. In this article, we also consider the intent of actors and not only the direct or indirect unintended consequences of AI use.

Given that the concepts mentioned above fall short in providing an adequate definition of the types of activities under analysis, we use the concepts of "malicious use and abuse" of AI, as proposed by [16]. By "malicious use" [14], [16], [27], [28] we refer to the use of AI to enhance, augment, or enable acts committed by individuals or organizations. This includes practices not necessarily considered crimes by specific legislation, but that still compromise the safety and security of individuals, organizations, and public institutions. By "malicious abuse" [16], we refer to the exploitation of AI with bad intentions, as well as attacks on AI systems themselves. Therefore, this study analyzes AI-enabled attacks (malicious use of AI) and the vulnerabilities of AI models (malicious abuse of AI).

V. OVERVIEW OF MALICIOUS USE AND ABUSE OF AI

After analyzing the academic literature (43 papers, books, and conference proceedings), reports (5 reports), and other documents (26 sources, including news stories, web pages, and other general documents), it was possible to identify the main malicious uses and abuses of AI systems.

A. MALICIOUS ABUSE OF AI: VULNERABILITIES OF AI MODELS

1) INTEGRITY ATTACKS

Machine learning (ML) has become more prevalent in recent years. This has created incentives for attackers to manipulate models (e.g., the software itself) or the underlying data, making ML models prone to integrity attacks. In integrity attacks, hackers attempt to inject false information into a system to corrupt the data, undermining their trustworthiness [29, p. 89]. One of the risks associated with the vulnerability of AI models is the creation of 'adversarial examples'. According to [30, p. 1], "adversarial examples are malicious inputs designed to fool machine learning models" which causes misclassification of material scrutinized by the systems. In some cases, the perturbations are too subtle to be perceived by human observers, but they still cause AI systems to make mistakes [31], [32].

One example of an adversarial ML is a 'poisoning attack'. The attacker influences the training data of the system to alter the results of a predictive model by injecting a few corrupted points in the training process [33, p. 19]. In other words, poisonous samples can be injected into the training data to manipulate the classifier, leading to undesirable consequences. A concrete example is the attack on Tay, Microsoft's AI chatbot, which was released in 2016. The chatbot had the objective of creating tweets that could not be distinguished from a human actor. Within a few hours of release, users launched a coordinated attack in which they tweeted offensive words and phrases, exploring Tay's "repeat after me" function. This led the bot to reproduce similarly objectionable content [34], [35]. According to [36], the Corporate Vice-President of Microsoft, "although we had prepared for many types of abuses of the system, we had made a critical oversight for this specific attack." Consequently, after less than 16 hours, Microsoft had to suspend the account. This demonstrates that defending a chatbot against attacks is challenging, especially when the system is trained in online environments with unforeseeable live interactions [37, p. 103].

Researchers at New York University (NYU) explored another risk associated with the context of outsourced training data [38]. They demonstrated that an adversary might create a BadNet (a maliciously trained network), which displays conventional behavior until a potential attacker triggers an attack. To test this hypothesis, BadNets were implemented in a complex traffic sign detection system. They demonstrated that a stop sign could be correctly identified by a self-driving car until a stop sign with a pre-defined trigger (yellow 'Post-It' note) was presented. This study demonstrates that AI models might be susceptible to data poisoning and adversarial examples, resulting in misclassifications and errors with potentially grave consequences that are difficult to foresee for humans unfamiliar with the technology. This might be one of the reasons why the recently proposed EU AI Act entails specific requirements for training data of 'high-risk systems' in Article 10 [21, pp. 48–49].

2) UNINTENDED OUTCOMES OF THE USE OF AI

Models used to train AI systems can present a different result from what was expected by the developer for various reasons. For instance, models based on neural networks may unintentionally memorize and disclose details. This can be problematic, especially when the data used to train the models are private or sensitive. [39] explained the phenomenon: during the learning process, such models might memorize details unrelated to the primary task. To prevent harmful consequences from unintended memorization and disclosure of information by the algorithm, it is necessary to apply techniques that guarantee data privacy.

The team behind the development of Smart Compose, the real-time suggestion system used by Google's Gmail service, considered this carefully [40]. To avoid unintended memorization, they conducted ''extensive testing to make sure that only common phrases used by multiple users are memorized'' [40, p. 2294]. Their goal was to prevent the models from learning details (e.g., private information) that were not related to the primary task (e.g. general and commonly used phrases) while training the algorithm. For example, when a user enters a text prefix such as ''my ID number is'', the model should not suggest a text completion with the ID number of another user see [39]). This challenge serves as one example in which the developer does not have the malicious intent of disclosing the user's personal information; the potential harm resides in the possibility that the model performs differently than previously expected (i.e., by memorizing private data).

3) ALGORITHMIC TRADING/STOCK MARKET MANIPULATION

With the help of computers and AI-powered software programs, technology facilitates and accelerates the pace of financial analysis and decisions. The use of AI systems in market trading, which causes it to move ''with lightning speed'' [41, p. 411], has both positive and negative aspects. In terms of positive aspects, the current financial technology has, for instance, decreased transactional charges and costs of capital for businesses [42, p. 1273]. However, algorithmic trading with decisions that are difficult to follow for humans inserts instability into the market. As a result, a risk for high-speed crashes (i.e., flash crashes) emerges. David Weild IV, the former vice-chairperson of Nasdaq, bluntly argued that ''we've created a stock market that moves too darn fast for human beings'', which is the reason ''we see shocking results'' [43].

The challenges of automated decision-making in the financial sector became apparent after the 2010 flash crash, which caused a loss of almost $1 trillion. Navinder Singh Sarao, a high-frequency trader, was sentenced in 2020 to a year of home incarceration for his involvement in this incident [44]. Sarao was accused of using an automated program to create large sell orders to push down prices [45]. Once the prices dropped, he canceled orders to buy at lower market prices to get the benefits when the market recovered. This first market crash in the era of algorithmic trading served as ''a wake-up call'' [46] not only to traders but also to regulators, showing some of the challenges of high-speed automated trading and automated-decision making more generally. To prevent similar incidents in the future, some techniques used to manipulate high-frequency trading, such as spoofing and layering, were banned [47].

The discussions surrounding the development of regulatory frameworks usually focus on market harm caused by malicious actors [41, p. 412]. Even though this is a necessary evaluation, it is also important to consider what could be done in the case of a technological accident or insufficient testing. As trading on stock markets becomes increasingly driven by algorithms, investors could face similar flash crashes more often. In such an environment, things can change and ''get out of hand in seconds'' [48]. Among the potential policy responses to flash crashes is the creation of insurance systems. [49, pp. 1094–1095] suggests that a financial market fund named the ''National Protection Fund'', which would compensate the investor eventually harmed by market disruptions caused by algorithms, could be a way of guaranteeing more stability and safety in trading. In addition, strengthening cybersecurity and an in-depth assessment of the respective algorithms could help to prevent the harmful consequences of high-speed crashes.

4) MEMBERSHIP INFERENCE ATTACKS

In membership inference attacks, the malicious actor aims to uncover and reconstruct the samples used to train a ML model [50, p. 3]. These attacks can be effective on several systems, such as classification and sequence-to-sequence models [51]. They can also be used against generative adversarial networks (GANs). GANs are a class of deep-learning model that creates seemingly realistic - but fake - examples of the data used in the training process. This technique is used in different applications, such as the website https://thispersondoesnotexist.com/.

In a recent study, [50] demonstrated that the faces produced by the ''This person does not exist'' algorithm are quite similar to the faces of the individuals that were part of the training data. The authors of the study concluded that, through membership inference attacks, it is possible to identify samples that are not identical, but that share the same identity. This could enable attackers to discover the real face of the people whose photos were part of the training datasets.

Therefore, membership inference attacks have privacy ramifications, affecting the individuals whose faces were used to train ML models. For instance, if a similar attack was used on a medical data model, attackers might be able to link a disease to an existing person [51, p. 1]. Such attacks are not limited to models using datasets of biometric data (e.g. images of faces, voice recordings, gait detection) but could also include others built on highly sensitive information such as genetic data. Potential venues to mitigate the risk of membership inference attacks include to ensure that models are being trained on diverse datasets, reduce dataset bias, as well as conducting extensive prior testing to ensure the system is not prone to such an attack.

B. MALICIOUS USE OF AI: AI-ENABLED AND AI-ENHANCED ATTACKS

1) SOCIAL ENGINEERING

Social engineering attacks use deception techniques to manipulate human subjects to share sensitive or personal information, which can be used for fraudulent purposes [52]. Such attacks are performed in different ways using an array

TABLE 1. Summary of malicious abuse of AI.

Integrity Attacks	Adversarial examples, a type of integrity attack, can be used to manipulate ML models, causing the algorithm to make mistakes (e.g. misclassification) [32]. Example: Microsoft's Tay.
Unintended AI Outcomes	Algorithms can present an unexpected output due to, for instance, unintentional memorization by models based on neural networks [39]. Example: Gmail's Smart Compose.
Algorithmic Trading	With the increase of algorithmic trading, the stock market is susceptible to high-speed crashes [41]. The incidents can be intentional or accidental. Example: 2010 Flash Crash.
Membership Inference Attacks	Such attacks try to uncover and reconstruct data used to train Machine Learning models [50]. The attacks can target datasets containing, for instance, biometric and genetic data. Example: thispersondoesnotexist.com.

of AI techniques. Using these techniques, cybercriminals can create elegant manipulation tactics, consequently increasing their chances of success and gains.

1.1) Deception and Phishing

Hackers can use AI techniques to develop a 'social bot', which can help them deceive and manipulate a person into complying with their request [53]. These 'social bots' are algorithms designed to emulate human behavior by producing content and interacting with users on the internet [54, p. 96], [55, pp. 556–557]. For instance, the request of social bots can access a website that enables the criminal to take over the computer of the victim. One of the first known cyberattacks that used AI techniques was a dating chatbot known as 'CyberLover' [56]. It was released in 2007 to lure users of chat rooms into sharing personal information or click on fraudulent links. The bot used natural language processing (NLP) to deliver a customized dialog, which raised concerns about the capabilities being used in cybercrime.

Similarly, attackers can masquerade themselves as trusted individuals or companies to induce the victim to open an email or link to steal data. The technique, known as phishing, can also be enhanced by AI to maximize the reach and gain of criminals. This was demonstrated by [57], who conducted an experiment using a model based on machine learning techniques to generate text to be posted on Twitter. The authors chose this social media platform because of the character limitation of each tweet, which makes posts with broken English and shortened links to be considered acceptable and normal. The results show that the dynamics of such platforms may facilitate the use of machine-generated text for phishing. AI may enable growth in these types of attacks in social media because posts tend to be written in an informal tone, with occasional spelling and grammar mistakes, and with shortened links.

1.2) Big Nudging and Manipulation

In addition to the potential targeted action described in the previous section, large numbers of bots might be created to support actions with malicious intent. Bots can potentially influence public opinion and the outcome of elections [58], [59]. For instance, by retweeting specific content or replicating hashtags, social bots can be used to create the impression that a candidate or political movement is more popular, deceiving users on social media platforms. A similar strategy is astroturfing, a process that mimics a bottom-up activity to create the impression that a policy or individual has widespread grassroots support when little or no support exists [60], [61]. An example of this is when a given organization is responsible for publishing thousands of Twitter posts using different accounts to influence public opinion against or in favor of a candidate in an election see [62]). Astroturfing can be found in Twitter posts, blogs, news portals, and other online platforms, and they can be used as disinformation strategies [62], [63].

Bots can also be used to create the perception of support for a cause in public consultations and interfere with polls. Concerns over this possibility spiked after the Federal Communications Commission's (FCC) consultation on net neutrality in the United States [64]. As the FCC had plans to roll back net neutrality protections, the regulator opened a consultation to gather public opinion on the topic through a comment section. The data analytics company Gravwell identified that, out of the approximately 22 million comments received by FCC, more than 80% were submitted by bots [65]. In this case, natural language generation was used to artificially inflate the support against net neutrality protection.

Another use of AI in this context is online profiling and targeting. The Cambridge Analytica scandal exemplifies this. According to reports and whistle-blowers, the app GSRApp was used to deceptively collect the personal data of their users, including personality traits, which were later used to train an algorithm [66, p. 7]. This algorithm generated personality scores for app users and their Facebook friends, which were then matched with the US elector records. Cambridge Analytica used the resulting data to develop voter profiling and targeted advertising services [66, p. 7]. With such information, politics could target specific groups of people by manipulating messages tailored to their psychological profile, in addition to disinformation and inflammatory material. Using these tools to change the behavior of individuals through manipulation can impact democratic processes and election outcomes.

2) MISINFORMATION AND FAKE NEWS

The development and diffusion of technology, blogging platforms, and social media have changed the way individuals consume information, access news items, and form opinions. The fast pace of the Internet also enables anyone to create and rapidly share content, which can reach many people. This scenario has created an environment that allows the creation and spread of misinformation and fake news. Although the term ''fake news'' is contested by some journalists and academics [67]–[69], it is still relevant to promote debates on digital literacy and encourage scholarly work on the issue [70]. Moreover, the justification behind the call for a ban has been demonstrated to be insufficient for abandoning the term [71].

Unsubstantiated rumors, speculation, and deliberately false information can lead to disastrous consequences, especially in times of uncertainty and social unrest, such as endemics and pandemics [72], [73]. During political events such as elections, it can also be harmful [74]–[76]. AI systems can fuel the creation and spread of this type of content, which represents a risk to society and democratic processes, potentially even democracy as such [77].

Tools such as GPT-3 could boost the creation of written pieces aimed at misinformation. GPT-3 is an autoregressive language model that uses deep learning to complete tasks such as question-answering, text completion, and summarization [78, p. 681], [79, p. 1]. Due to format, choice of words, and consistency, texts created automatically with the tool might look like they were written by a human, misleading the reader due to apparent credibility [79]. Some examples of this can be seen on the website ''NotRealNews.net'' [80], which uses AI to generate AI-written fake news pieces. The idea behind the project was to demonstrate how this tool can be used to support the work of journalists. Considering that the articles were mostly convincing, such a tool could easily be used to disseminate compelling fake news articles. This means that automatically generated texts, coupled with current targeting capabilities, could further increase the quantity, quality, and impact of fake news and disinformation campaigns. These might impact democratic processes to a greater (e.g., by convincing electors to change their vote) or to a lesser degree (e.g., by confirming or reinforcing electors' pre-existing views) [81, p. 977]. In addition, as technology evolves, texts can be tailored to the audience's taste, increasing the proliferation of ''filter bubbles'' and polarization [78, p. 692].

Some strategies could help to reduce the negative impact of the use of AI systems to create and disseminate fake news and misinformation. [82] conducted a study that revealed that information literacy increases the likelihood of identifying fake news pieces. According to the Association for College and Research Libraries (ACRL), information literacy is ''the set of integrated abilities encompassing the reflective discovery of information, the understanding of how information is produced and valued, and the use of information in creating new knowledge and participating ethically in communities of learning'' [83, p. 8]. For this reason, educating individuals about the adequate use of digital resources is of paramount importance. Following this logic, the more citizens can navigate the online environment and critically evaluate the information, the less unfounded stories will impact them and their community [82, pp. 13–14].

In addition, information systems and providers play important roles. Given that many users access news and information based on algorithmic decisions, social media platforms (e.g. Facebook) and search engines (e.g. Google) have been facing pressure to revise and improve their algorithms to structure the presentation of content differently (e.g. in the context of the debate on a 'right to be forgotten' [84]), as well as to reduce the quantity of fake news appearing on their feeds [85], [86]. Platforms are not mere intermediates; their algorithms are designed to deliver specific types of content to users based on past activities and foster user engagement. In this model, a person who reads one or more pieces from a news outlet that disseminates false information is likely to receive additional content from the same source, since it creates engagement for the platform. If tech companies proactively enhance their algorithms, the detection of unreliable sources can be improved, and their spread contained. At the same time, more human monitoring and oversight are indispensable, since only this type of control is capable of understanding information in context. There seems to be some development in these areas – for instance, with the development of a somewhat independent oversight board by Facebook (now Meta) to ''oversee'' and check important decisions, although much still remains to be discovered and done [87], [88].

Finally, it is worth mentioning that civil society organizations and activists can make positive contributions. An example is the initiative ''Sleeping Giants Brasil'', which was inspired by the Twitter account created in the United States after the 2016 election, called ''Sleeping Giants'' [89]. Using a Twitter account, activists planned to reduce the advertising revenue of certain news outlets known to spread disinformation. The revenue is the product of the affiliation of these websites with Google's ad platform. Websites that share fake news stories earn money according to the number of views and clicks on ads displayed on websites. After taking screenshots of ads on such websites, the account would publicly question brands about their support of that type of content. Many companies would then block their ads from appearing on such websites in the future, consequently reducing their stream of income. Many similar accounts were created in Brazil to combat the spread of fake news.

3) HACKING

3.1 FORGERY: DEEPFAKES

Prominent examples of forgery in the digital age are deepfake videos and images. Such hyper-realistic media may apply AI in its creation to portray a person saying or doing things that did not happen. [90], [91]. The use of AI for the forgery of videos and images enables more realistic material, making it difficult to distinguish between what is real and what is fake. Although such manipulation is not new, especially after the popularization of programs such as Photoshop, AI makes forgery more elaborate and challenging to detect. For instance, Ali Aliev developed a method for creating deepfakes in real time [92]. To test the tool, the programmer joined a random Zoom meeting pretending to be Elon Musk.

This example goes along with the current practice of mostly using the figures of well-known individuals, such as celebrities and politicians, in deepfake materials [93]. The danger of these videos and images resides in the fact that they can be created for several malicious purposes: propaganda, disinformation, bullying, revenge porn, or blackmail to name just a few [94].

The malicious use of forged videos can have a direct impact on politics and international relations. The Democratic Party in the United States created a fake video of the chairman at a convention to highlight their concern for the effect of deepfakes in democratic processes [95]. One of the alternatives to reduce the negative consequences of the use of forged videos is to raise awareness of the population about such technology use. Bruno Sartori, a deepfake creator, produces humorous videos depicting Brazilian national politics, especially involving politicians from the executive branch. Adding a level of absurdity in the videos, viewers understand that they are not real and the material produced constitutes an elaborate satire [96]. More importantly, the material shared on social media platforms serves to demonstrate the risks of the technology to the public. Inoculation theory helps explain such interventions [97], [98]. According to this theory, prior exposure can help protect individuals against future threats. In the context of deepfakes, by offering knowledge about the technology and convincing the population to interpret videos critically, such initiatives might help individuals to be ''inoculated'' against maliciously forged videos. In addition to raising awareness, it is important to further develop tools for deepfake detection. AI techniques can be particularly helpful, such as the use of recurrent neural networks [99].

[100] describe a phenomenon known as the ''liar's dividend'', which adds a layer of complexity to the problem. According to the authors, liar's dividend refers to the situation in which someone, a 'liar', takes advantage of the existence of deepfake videos to discredit a real video. This person would claim that the material was manipulated, creating doubt about its authenticity among the public. The more the public is aware of the use of AI to doctor videos, the more skeptical they will be, questioning videos and images that are, in fact, real. This is what the authors called the liar's dividend: ''this dividend flows, perversely, in proportion to success in educating the public about the dangers of deep fakes'' [100, p. 1785]. Therefore, there is a possibility that, during elections, a candidate that was caught on tape might lie about the video, saying it is a deepfake, convincing electors of their innocence. At this point, it remains to be seen whether and how regulations such as the EU AI Act will be able to address deepfakes. In the current draft, no dedicated prohibition is visible. However, the People's Republic of China is introducing relevant legislation that will require platform operators to prevent the spread of deepfakes on their networks [101].

3.2 REPETITIVE TASKS

AI is also efficient in conducting repetitive tasks that can be used maliciously. One example is the incident involving the company Ticketmaster. AI tools were employed to bypass Captcha,[3] which enabled the purchase of thousands of tickets that would later be resold to generate profit [102]. Pattern recognition is not a problem limited to Captcha-defeating purposes [47]. Concerns about other hacking-based crimes, such as password-cracking, should also be considered. One way to crack passwords is through brute force attacks, which can be time and resource consuming. However, it has been demonstrated that brute-force attacks using AI have a significantly higher success rate than non-AI based attacks [103]. In other words, the advances in AI could lead to repetitive tasks being used for malicious purposes, such as password cracking.

3.3 MALWARE

Malware threats have been used for several decades. Creeper Worm, the first documented malicious software, appeared in the 1970s [104]. Since then, these attacks have become a massive industry that is now a significant cybersecurity concern. The AV-TEST Institute registers more than 350,000 new malware and potentially unwanted applications (PUA) per day [105]. This means that four new malware or PUAs are registered every second. As malware developers continue to innovate and create more elaborate malicious programs, it becomes challenging to establish proper and timely defense mechanisms. Currently, concerns revolve around the possibility of AI techniques being used to create more effective and difficult to detect malware [16], [47]. However, to the best of our knowledge, this technology is not yet well developed.

The current possibilities are mainly explored by academic research and as proof of concept by companies. For instance, IBM presented DeepLocker at the Black Hat USA 2018 [106]. This system enhances malware with AI and improves its evasion capabilities. DeepLocker explores the lack of explicability of AI systems, which is mainly considered a weakness of AI, to its advantage [16, p. 8]. It uses a deep neural network to select targets and conceal the intent until it reaches the desired destination. The main risk of this type of AI-enhanced malware is that it can infect many systems without being detected. In addition, the capabilities of developing systems such as DeepLocker are not constrained to states; civilians and private organizations can also work on the development of such high-risk malware [107]. Thus, even if AI-enabled or AI-enhanced malware are not well developed now, the potential risks associated with such a possibility need to be considered.

One way of addressing the challenges of AI-based or AI-enhanced malware is to improve capabilities in the field of cyber autonomy. The feasibility of cyber autonomy was demonstrated during the Cyber Grand Challenge, hosted by the Defense Advanced Research Projects Agency (DARPA) in 2016. The finalist teams of the competition were asked

[3] CAPTCHA stands for Completely Automated Public Turing Test to Tell Computers and Humans Apart. One of its purposes is to prevent bots from accessing certain content on websites to avoid malicious attacks.

to "develop automated cyber defense systems that can self-discover, prove, and correct software vulnerabilities at real-time" [108, p. 173]. During the competition, the systems were able to auto-detect and correct. In addition, they were able to attack the software of other participants in their network. According to [108, p. 173], since this event, it was possible to identify a movement towards "security automation". This can be considered the first step toward cyber autonomy. Developing capabilities in autonomous defensive cybersecurity is a way of leveraging AI systems against malicious actors. However, given the dual-use property of the technology, software created for defense can also be used for offensive purposes. To reduce this risk, there needs to be clear regulations around these systems' use and security safeguards.

4) AUTONOMOUS WEAPONS SYSTEMS (AWS)

Militaries have been exploring the possibility of autonomy in weapons for some time, practically since the inception of AI in the late 1950ies [109, pp. 8–11]. As machines can process data, analyze information, and make decisions in some situations in less time than humans, their use is particularly attractive in the context of defense. While Autonomous Weapons Systems (AWS) promise military and strategic advantages [110], [111], they also come with risks [112]. AWS can be defined as AI systems designed to select (i.e., search for or detect) and engage (i.e., use force against) targets without the need for human control or human action after its activation [113, pp. 13–14], [114, p. 1]. Autonomous functions can be applied to different platforms, such as ships or fighter jets.

One of the risks of this emerging technology is the possibility of the software embedded in military hardware (e.g., drones) being altered by malicious actors. If a drone is hacked and the GPS location of an attack changed, it would behave according to the new rules set in the software. This could result in unintended casualties due to the target being redirected. Similarly, if the data used to train the systems are poisoned, this could lead to disastrous consequences. In 2014, Reprieve published a report demonstrating that drone attacks aimed at killing 41 individuals resulted in the death of approximately 1,147 people, raising questions about the accuracy and precision of 'targeted killing' [115]. Gibson, who led the report, argued that drone strikes are "only as precise as the intelligence that feeds them" [116]. Such high risks associated with attacks on AI systems used in warfare are being discussed in academia [2], [110], [117], [118], civil society [119], [120], and at the government level [121]. However, at present, there are no international regulations regarding the use of AWS.

The implications of the use of AI in warfare were first debated among state parties to the United Nations Convention on Certain Conventional Weapons (CCW). The main purpose of the CCW is "to ban or restrict the use of specific types of weapons that are considered to cause unnecessary or unjustifiable suffering to combatants or to affect civilians indiscriminately" [122]. Within the CCW, the topic is mainly discussed through the lens of international humanitarian law. Ethical issues, for instance, play a secondary role. From 2014 to 2016, annual Informal Meetings of Experts on AWS were held in Geneva. Later, the CCW created a Group of Governmental Experts (GGE) on AWS which is the main forum for debating autonomous weapons systems at the international level [123, pp. 188–189].

TABLE 2. Summary of malicious use of AI.

Deception and Phishing (Social Engineering)	To develop social bots, attackers can use AI techniques, such as natural language processing [53]. The bots are used to deceive and manipulate people into complying with their requests (e.g., sharing personal information). Example: Cyberlover.
Manipulation (Social Engineering)	Malicious actors can use AI techniques to develop algorithms or social bots to manipulate public opinion. Example: Cambridge Analytica.
Misinformation and Fake News	AI systems can be used to accelerate the creation and spread of unsubstantiated content aimed at misinformation [78, p. 692]. Example: tools such as GPT-3.
Deepfakes (Hacking)	With the advances in AI, algorithms support the creation of hyper-realistic images and videos, known as deepfakes [91]. Example: "fake" Elon Musk joining Zoom meeting.
Repetitive Tasks (Hacking)	AI systems can perform repetitive tasks efficiently, which malicious actors can exploit [102], [103]. Example: Captcha-defeating and password cracking.
Malware (Hacking)	Malware could be enhanced with AI techniques, improving its capabilities. Currently, the possibilities are investigated by academic research and as proof of concept [107]. Example: DeepLocker.
Autonomous Weapons Systems (AWS)	Malicious actors could alter the software embedded in weapons systems, resulting in unintended casualties. Example: hacking the GPS of drones.

Among the possibilities for regulation is the creation of an additional protocol to the existing convention. This would follow previously adopted additional protocols, such as those involving weapons with non-detectable fragments, landmines, incendiary weapons, blinding laser weapons, and explosive remnants of war. However, in the past, negotiations that started in the CCW, such as the one on cluster munitions, were moved outside the CCW due to a lack of consensus. In the case of cluster munitions, some of the CCW's treaty members started negotiations outside the CCW in February 2007 [124]. As a result, the Cluster Munitions Convention was adopted in May 2008 and has 110 state parties as of August 2021 [125]. The treaty was made among states that were initially in agreement and later adopted by others as well, which might be the way forward with AWS.

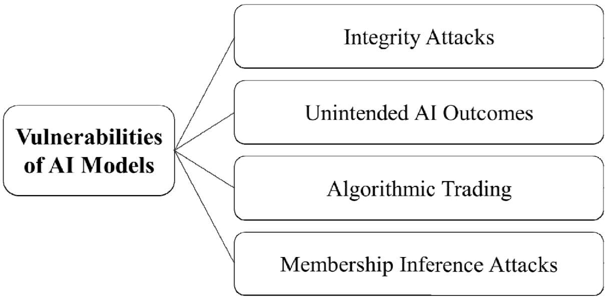

FIGURE 1. Malicious Abuse of AI.

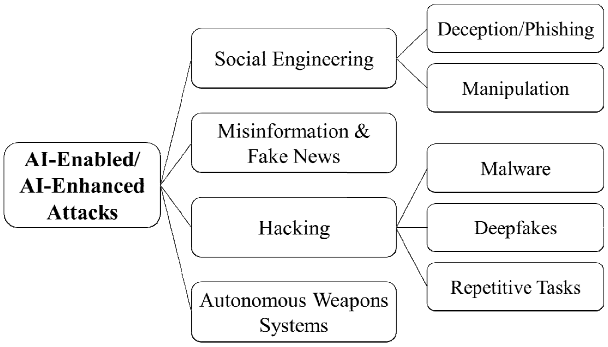

FIGURE 2. Malicious Use of AI.

VI. RESULTS

In the previous sections, we illustrated AI misuse through the lens of our definition of malicious use and abuse of AI. This results in a typology that distinguishes between different types (summarized in Tables 1 and 2). The category 'malicious abuse of AI' (Figure 1) encompasses the exploitation of AI vulnerabilities, be it via integrity attacks on either the learning models or the learning data. Furthermore, we include unintended AI outcomes (such as Google's Smart Compose) - albeit falling outside of our focus on intentional AI crime - as it has the potential for intentional exploitation. Finally, we included algorithmic trading and membership inference attacks.

Within the category of 'malicious use of AI' (Figure 2) fall AI enabled and AI enhanced attacks on both physical (e.g., human) and digital targets (e.g., data infrastructures and computer systems). Such attacks can be further subdivided into four categories: (1) social engineering, (2) hacking, (3) misinformation and fake news, and (4) AWS (see Table 2 for a summary).

The resulting typology is comprehensive, but by no means final or complete. Certain categories overlap, and others may emerge in the future as technology evolves. However, this structured overview of the current state of the art and the different attack vectors provides a useful overview over the emerging field of AI crime.

VII. DISCUSSION

AI techniques are increasingly deployed in different areas and for an increasing number of purposes. This brings both benefits and risks to society [14]. Among the risks are the use and abuse of AI systems with malicious intent. Even though the capabilities of AI-enhanced technology might not always lead to more sophisticated attacks, they certainly have the potential to increase scale and reach. Cybercriminals will progressively integrate AI techniques and the use of AI systems in their plans.

The risks presented in our overview are especially challenging when cybercriminals exploit systems during periods of societal instability. This is facilitated during the COVID-19 pandemic, which caused a growth in the number of people using online tools to work and socialize. The massive shift of social interaction to the online environment increased security vulnerabilities, which malicious actors already exploit at an alarming rate [126]. Not only were individuals and small businesses targeted; in fact, Interpol identified that cybercriminals focused on critical infrastructure, major corporations, and governments [127]. Given the potential impacts of such attacks, it is vital to consider and mitigate these risks.

Some of the issues presented in this overview have been discussed elsewhere [15], [16]. However, in addition to adding novel types of threats in our typology (e.g. Membership Inference Attacks) and providing salient examples, we also provided a different classification than previous works. We divide the attacks between (1) AI-Enabled/ AI-Enhanced attacks and (2) vulnerabilities of AI models. We submit that such separation is helpful because different strategies can alleviate the risks.

Addressing challenges linked to vulnerabilities of AI models is highly dependent on the work of engineers and development teams. Developing robust AI systems is paramount. To this end, teams behind the development of algorithms should adhere to principles such as privacy-by-design. Organizations, government bodies, and scholars are developing and fine-tuning impact assessment tools for AI systems [128]–[130]. Such tools help translate relevant principles (such as privacy, transparency and fairness [131]) into practical evaluations. Efforts to identify risks via impact assessments are already conducted for data protection compliance in many countries, and similar initiatives can be helpful to deal with the challenges presented by AI systems.

When discussing ways of dealing with the risks presented by AI-Enabled/AI-Enhanced attacks, more is needed in prevention/proactive measures and adequate response. Given that regulatory frameworks and governance mechanisms might not be formulated at the same pace of technological advancements, it is vital to act proactively to reduce the risks outlined in this paper. Instead of finding one overarching solution, different sectors of society could gradually identify initiatives that can help build more resilience and preparedness. Initiatives with local communities, such as promoting data and information literacy, reducing digital divide gaps, and creating campaigns to raise awareness on AI-related threats can be a staring point.

Finally, we wish to emphasize that when discussing the challenges posed by AI systems, one should not forget that the possibilities are also limited. Some simple and easy tasks for humans (e.g., sensorimotor skills such as developing motor abilities through the senses) can be difficult or even impossible for computers to carry out. At the same time, some functions that are complex to humans can be quickly developed in AI systems (e.g., finding patterns in an extensive data set). This is the basis of what became known as Moravec's paradox: "it is comparatively easy to make computers exhibit adult-level performance in solving problems on intelligence tests of playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility" [132, p. 15]. Understanding the actual capabilities and limitations of emerging technologies such as AI is therefore critical for developing effective policies and strategies for living in a safer world.

VIII. CONCLUSION

The threats posed by the use and abuse of AI systems must be well understood to create mechanisms that protect society and critical infrastructures from attacks. Based on the available literature, reports, and previous incidents, we focused on creating a classification of how AI systems can be used or abused by malicious actors. This includes, but is not limited to, physical, psychological, political, and economic harm. We explored the vulnerabilities of AI models, such as unintended outcomes, and AI-enabled and AI-enhanced attacks, such as forgery. This article also describes past incidents, such as the 2010 flash crash and the Cambridge Analytica scandal, manifesting the challenges at hand. We also outlined attacks that, to the best of our knowledge, have only been demonstrated through "proof of concept", such as IBM's DeepLocker. In response to the risks presented in this paper, we have also explored some possible mitigation strategies. Industries, governments, civil society, and individuals should cooperate in developing knowledge and raising awareness while developing technical and operational systems and procedures to address the challenges.

Although this type of classification is a useful starting point, it does not come without drawbacks. Some AI-enabled or AI-enhanced attacks might not fit the categories established. Further work could use empirical methods to assess whether the classification scheme presented is generalizable and representative. When sufficient data is available, methods such as statistical analysis could be helpful to reach a more complete overview of the threat scenario. Continuously mapping the risks associated with malicious use and abuse of AI helps to enhance preparedness and increases the potential to prevent and adequately respond to attacks.

ACKNOWLEDGMENT

The authors would like to thank Alex Belloir for reviewing the manuscript.

REFERENCES

[1] K. Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. London, U.K.: Yale Univ. Press, 2021.

[2] D. Garcia, "Lethal artificial intelligence and change: The future of international peace and security," *Int. Stud. Rev.*, vol. 20, no. 2, pp. 334–341, Jun. 2018, doi: 10.1093/isr/viy029.

[3] T. Yigitcanlar, K. Desouza, L. Butler, and F. Roozkhosh, "Contributions and risks of artificial intelligence (AI) in building smarter cities: Insights from a systematic review of the literature," *Energies*, vol. 13, no. 6, p. 1473, Mar. 2020, doi: 10.3390/en13061473.

[4] I. van Engelshoven. (Oct. 18, 2019). *Speech by Minister Van Engelshoven on Artificial Intelligence at UNESCO, on October the 18th in Paris*. Government of The Netherlands. Accessed: Apr. 15, 2021. [Online]. Available: https://www.government.nl/documents/speeches/2019/10/18/speech-by-minister-van-engelshoven-on-artificial-intelligence-at-unesco

[5] O. Osoba and W. Welser IV, *The Risks of Artificial Intelligence to Security and the Future of Work*. Santa Monica, CA, USA: RAND Corporation, 2017, doi: 10.7249/PE237.

[6] D. Patel, Y. Shah, N. Thakkar, K. Shah, and M. Shah, "Implementation of artificial intelligence techniques for cancer detection," *Augmented Hum. Res.*, vol. 5, no. 1, Dec. 2020, doi: 10.1007/s41133-019-0024-3.

[7] A. Rodríguez-Ruiz, E. Krupinski, J.-J. Mordang, K. Schilling, S. H. Heywang-Köbrunner, I. Sechopoulos, and R. M. Mann, "Detection of breast cancer with mammography: Effect of an artificial intelligence support system," *Radiology*, vol. 290, no. 2, pp. 305–314, Feb. 2019, doi: 10.1148/radiol.2018181371.

[8] J. Furman and R. Seamans, "AI and the economy," Nat. Bur. Econ. Res., NBER, Cambridge, MA, USA, Work. Paper, 2018, doi: 10.3386/w24689.

[9] D. R. Coats, *Worldwide Threat Assessment of the U.S. Intelligence Community*. New York, NY, USA, 2017, p. 32.

[10] L. Floridi, "Soft ethics: Its application to the general data protection regulation and its dual advantage," *Philosophy Technol.*, vol. 31, no. 2, pp. 163–167, Jun. 2018, doi: 10.1007/s13347-018-0315-5.

[11] P. S. Chauhan and N. Kshetri, "2021 state of the practice in data privacy and security," *Computer*, vol. 54, no. 8, pp. 125–132, Aug. 2021, doi: 10.1109/MC.2021.3083916.

[12] S. Gordon and R. Ford, "On the definition and classification of cybercrime," *J. Comput. Virol.*, vol. 2, no. 1, pp. 13–20, Aug. 2006, doi: 10.1007/s11416-006-0015-z.

[13] *Cybercrime*. United Nations: Office Drugs. Accessed: May 19, 2021. http://www.unodc.org/unodc/en/cybercrime/index.html

[14] M. Brundage, S. Avin, J. Clark, and H. Toner, "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," 2018, *arXiv:1802.07228*.

[15] T. C. King, N. Aggarwal, M. Taddeo, and L. Floridi, "Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions," *Sci. Eng. Ethics*, vol. 26, no. 1, pp. 89–120, Feb. 2020, doi: 10.1007/s11948-018-00081-0.

[16] V. Ciancaglini, "Malicious uses and abuses of artificial intelligence," in *Trend Micro Research; United Nations Interregional Crime and Justice Research Institute (UNICRI); Europol's European Cybercrime Centre (EC3)*, Nov. 2020. [Online]. Available: https://www.europol.europa.eu/publications-documents/malicious-uses-and-abuses-of-artificial-intelligence

[17] K. D. Fiedler, V. Grover, and J. T. C. Teng, "An empirically derived taxonomy of information technology structure and its relationship to organizational structure," *J Manage. Inf. Syst.*, vol. 13, pp. 9–34, Jun. 1996, doi: 10.1080/07421222.1996.11518110.

[18] N. Bostrom, "Information hazards: A typology of potential harms from knowledge," *Rev. Contemp. Philosophy*, vol. 10, pp. 44–79, May 2011.

[19] W. B. Carper and W. E. Snizek, "The nature and types of organizational taxonomies: An overview," *Acad. Manage. Rev.*, vol. 5, no. 1, pp. 65–75, Jan. 1980.

[20] (Apr. 21, 2021). *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence—Artificial Intelligence Act*. European Commission. Accessed: May 19, 2021. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence

[21] *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence—Artificial Intelligence Act—Annexes to the Proposal*. European Commission. Accessed: May 19, 2021. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence

[22] N. Akdemir and C. J. Lawless, "Exploring the human factor in cyber-enabled and cyber-dependent crime victimisation: A lifestyle routine activities approach," *Internet Res.*, vol. 30, no. 6, pp. 1665–1687, Jun. 2020, doi: 10.1108/INTR-10-2019-0400.

[23] P. N. Grabosky, "Virtual criminality: Old wine in new bottles?" *Social Legal Stud.*, vol. 10, no. 2, pp. 243–249, Jun. 2001, doi: 10.1177/a017405.

[24] B. Hibbard, *Ethical Artificial Intelligence*, 1st ed. Madison, WI, USA, 2015.

[25] D. G. Johnson and M. Verdicchio, "Reframing AI discourse," *Minds Mach.*, vol. 27, no. 4, pp. 575–590, Dec. 2017, doi: 10.1007/s11023-017-9417-6.

[26] R. V. Yampolskiy, "Taxonomy of pathways to dangerous AI," Phoenix, AZ, USA, Tech. Rep., Feb. 2016, pp. 143–148.

[27] A. Guterres. (May 2020). *Protection of Civilians in Armed Conflict*. United Nations, S/2020/366. Accessed: Jun. 2, 2020. [Online]. Available: https://undocs.org/en/S/2020/366

[28] E. Zouave, T. Gustafsson, M. Bruce, K. Colde, M. Jaitner, and I. Rodhe, "Artificially intelligent cyberattacks," Swedish Defence Research Agency, FOI, Tech. Rep. FOI-R–4947-SE, Mar. 2020.

[29] J. Luo, T. Hong, and S.-C. Fang, "Benchmarking robustness of load forecasting models under data integrity attacks," *Int. J. Forecasting*, vol. 34, no. 1, pp. 89–104, Jan. 2018, doi: 10.1016/j.ijforecast.2017.08.004.

[30] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," 2016, *arXiv:1611.01236*.

[31] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[32] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," 2016, *arXiv:1607.02533*.

[33] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, San Francisco, CA, USA, May 2018, pp. 19–35, doi: 10.1109/SP.2018.00057.

[34] O. Schwartz, "In 2016, Microsoft's racist chatbot revealed the dangers of online conversation," *IEEE Spectr.*, to be published. Accessed: Apr. 13, 2021. [Online]. Available: https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation

[35] T. Zemčík, "Failure of chatbot tay was evil, ugliness and uselessness in its nature or do we judge it through cognitive shortcuts and biases?" *AI Soc.*, vol. 36, no. 1, pp. 361–367, Mar. 2021, doi: 10.1007/s00146-020-01053-4.

[36] P. Lee. (Mar. 25, 2016). *Learning from Tay's Introduction*. Microsoft Blog. Accessed: Apr. 30, 2021. [Online]. Available: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

[37] N. Baracaldo, B. Chen, H. Ludwig, and J. A. Safavi, "Mitigating poisoning attacks on machine learning models: A data provenance based approach," in *Proc. 10th ACM Workshop Artif. Intell. Secur.*, Dallas, TX, USA, Nov. 2017, pp. 103–110, doi: 10.1145/3128572.3140450.

[38] T. Gu, B. Dolan-Gavitt, and S. Garg, "BadNets: Identifying vulnerabilities in the machine learning model supply chain," 2017, *arXiv:1708.06733*.

[39] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The secret sharer: Evaluating and testing unintended memorization in neural networks," 2018, *arXiv:1802.08232*.

[40] M. X. Chen, B. N. Lee, G. Bansal, Y. Cao, S. Zhang, J. Lu, J. Tsay, Y. Wang, A. M. Dai, Z. Chen, T. Sohn, and Y. Wu, "Gmail smart compose: Real-time assisted writing," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Anchorage AK USA, Jul. 2019, pp. 2287–2295, doi: 10.1145/3292500.3330723.

[41] G. Scopino, *Algo Bots and the Law: Technology, Automation, and the Regulation of Futures and Other Derivatives*. Cambridge, U.K.: Cambridge Univ. Press, 2020.

[42] T. C. W. Lin, "The new market manipulation," *Emory Law J.*, vol. 66, pp. 1253–1314, Jul. 2017.

[43] D. Wiener-Bronner. (Feb. 5, 2018). *How the Dow Fell 800 Points in 10 Minutes*. CNNMoney. Accessed: Jun. 24, 2021. [Online]. Available: https://money.cnn.com/2018/02/05/news/companies/dow-800-points-10-minutes/index.html

[44] K. Martin. (May 7, 2020). *Flash Crash—The Trading Savant Who Crashed the U.S. Stock Market*. Financial Times. Accessed: Apr. 14, 2021. [Online]. Available: https://www.ft.com/content/5ca93932-8de7-11ea-a8ec-961a33ba80aa

[45] S. N. Lynch and D. Miedema. (Apr. 22, 2015). *U.K. Speed Trader Arrested Over Role in 2010*. Flash Crash. Reuters, Washington, DC, USA. Accessed: Apr. 14, 2021. [Online]. Available: https://www.reuters.com/article/us-usa-security-fraud-idUSKBN0NC21220150422

[46] R. Wigglesworth. (Jan. 9, 2019). *Volatility: How 'algos' Changed Rhythm Market*. Financial Times. Accessed: Jun. 26, 2021. [Online]. Available: https://www-ft-com/content/fdc1c064-1142-11e9-a581-4ff78404524e

[47] A. Zwitter. (Jul. 27, 2017). *The Artificial Intelligence Arms Race*. Policy Forum. Accessed: Apr. 12, 2021. [Online]. Available: https://www.policyforum.net/artificial-intelligence-arms-race/

[48] J. Cox. (Feb. 16, 2018). *The Stock Market Correction Two Weeks Later: How it Happened, and if it Can Happen Again*. CNBC. Accessed: Jun. 24, 2021. [Online]. Available: https://www.cnbc.com/2018/02/16/the-stock-market-correction-two-weeks-later.html

[49] Y. Yadav, "The failure of liability in modern markets," *Virginia Law Rev.*, vol. 102, pp. 1031–1100, May 2016.

[50] R. Webster, J. Rabin, L. Simon, and F. Jurie, "This person (Probably) Exists. Identity membership attacks against GAN generated faces," 2021, *arXiv:2107.06018*.

[51] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership inference attacks on machine learning: A survey," 2021, *arXiv:2103.07853*.

[52] F. Mouton, L. Leenen, and H. S. Venter, "Social engineering attack examples, templates and scenarios," *Comput. Secur.*, vol. 59, pp. 186–209, Jun. 2016, doi: 10.1016/j.cose.2016.03.004.

[53] C. Freitas, F. Benevenuto, S. Ghosh, and A. Veloso, "Reverse engineering socialbot infiltration strategies in Twitter," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Paris, France, Aug. 2015, pp. 25–32, doi: 10.1145/2808797.2809292.

[54] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, "Design and analysis of a social botnet," *Comput. Netw.*, vol. 57, no. 2, pp. 556–578, Feb. 2013, doi: 10.1016/j.comnet.2012.06.006.

[55] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Commun. ACM*, vol. 59, no. 7, pp. 96–104, Jul. 2016, doi: 10.1145/2818717.

[56] S. Rossi. (Dec. 15, 2007). *Beware the CyberLover that Steals Personal Data*. PCWorld. Accessed: May 11, 2020. [Online]. Available: https://www.pcworld.com/article/140507/article.html

[57] J. Seymour and P. Tully. (2016). *Weaponizing Data Science for Social Engineering: Automated E2E Spear Phishing on Twitter*. [Online]. Available: https://www.blackhat.com/docs/us-16/materials/us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter-wp.pdf

[58] A. Bessi and E. Ferrara, "Social bots distort the 2016 U.S. Presidential election online discussion," *1st Monday*, vol. 21, no. 11, Nov. 2016.

[59] G. P. Nobre, J. M. Almeida, and C. H. G. Ferreira, "Caracterização de bots no Twitter durante as eleições presidenciais no Brasil em 2018," in *Anais do VIII Brazilian Workshop Social Netw. Anal. Mining (BraSNAM)*, Jul. 2019, pp. 107–118, doi: 10.5753/brasnam.2019.6553.

[60] M. Kovic, A. Rauchfleisch, M. Sele, and C. Caspar, "Digital astroturfing in politics: Definition, typology, and countermeasures," *Stud. Commun. Sci.*, vol. 18, no. 1, Nov. 2018, doi: 10.24434/j.scoms.2018.01.005.

[61] S. Mahbub, E. Pardede, A. S. M. Kayes, and W. Rahayu, "Controlling astroturfing on the Internet: A survey on detection techniques and research challenges," *Int. J. Web Grid Services*, vol. 15, no. 2, p. 139, 2019, doi: 10.1504/IJWGS.2019.099561.

[62] F. B. Keller, D. Schoch, S. Stier, and J. Yang, "Political astroturfing on Twitter: How to coordinate a disinformation campaign," *Political Commun.*, vol. 37, no. 2, pp. 256–280, Mar. 2020, doi: 10.1080/10584609.2019.1661888.

[63] A. Zwitter. (Jun. 12, 2016). *The Impact of Big Data of International Affairs*. Clingendael Spectator. Accessed: Apr. 7, 2021. [Online]. Available: https://spectator.clingendael.org/en/publication/impact-big-data-international-affairs

[64] I. Lapowsky. (Nov. 28, 2017). *How Bots Broke the FCC's Public Comment System*. Wired. Accessed: Apr. 8, 2021. [Online]. Available: https://www.wired.com/story/bots-broke-fcc-public-comment-system/

[65] C. Thuen. (Oct. 2, 2017). *Discovering Truth Through Lies on the Internet—FCC Comments Analyzed*. Gravwell. Accessed: Apr. 8, 2021. [Online]. Available: https://www.gravwell.io/blog/discovering-truth-through-lies-on-the-internet-fcc-comments-analyzed

[66] V. Bakir, "Psychological operations in digital political campaigns: Assessing Cambridge analytica's psychographic profiling and targeting," *Frontiers Commun.*, vol. 5, p. 67, Sep. 2020, doi: 10.3389/fcomm.2020.00067.

[67] J. Habgood-Coote, "Stop talking about fake news!" *Inquiry*, vol. 62, nos. 9–10, pp. 1033–1065, Nov. 2019, doi: 10.1080/0020174X.2018.1508363.

[68] M. Sullivan. (Jan. 8, 2017). *It's Time to Retire the Tainted Term*. Fake News, The Washington Post. Accessed: Apr. 29, 2021. [Online]. Available: https://www.washingtonpost.com/lifestyle/style/its-time-to-retire-the-tainted-term-fake-news/2017/01/06/a5a7516c-d375-11e6-945a-76f69a399dd5_story.html

[69] E. Zuckerman. (Jan. 31, 2017). *Stop Saying 'Fake News'. It's Not Helping*. Ethan Zuckerman. Accessed: Apr. 29, 2021. [Online]. Available: https://ethanzuckerman.com/2017/01/30/stop-saying-fake-news-its-not-helping/

[70] S. Alonso García, G. Gómez García, M. Sanz Prieto, A. J. Moreno Guerrero, and C. Rodríguez Jiménez, "The impact of term fake news on the scientific community. Scientific performance and mapping in web of science," *Social Sci.*, vol. 9, no. 5, p. 73, May 2020, doi: 10.3390/socsci9050073.

[71] J. Pepp, E. Michaelson, and R. Sterken, "Why we should keep talking about fake news," *Inquiry*, vol. 65, no. 4, pp. 471–487, Nov. 2019, doi: 10.1080/0020174X.2019.1685231.

[72] S. O. Oyeyemi, E. Gabarron, and R. Wynn, "Ebola, Twitter, and misinformation: A dangerous combination?" *BMJ*, vol. 349, pp. g6178–g6178, Oct. 2014, doi: 10.1136/bmj.g6178.

[73] J. Roozenbeek, C. R. Schneider, S. Dryhurst, J. Kerr, A. L. J. Freeman, G. Recchia, A. M. van der Bles, and S. van der Linden, "Susceptibility to misinformation about COVID-19 around the world," *Roy. Soc. Open Sci.*, vol. 7, no. 10, Oct. 2020, Art. no. 201199, doi: 10.1098/rsos.201199.

[74] W. L. Bennett and S. Livingston, "The disinformation order: Disruptive communication and the decline of democratic institutions," *Eur. J. Commun.*, vol. 33, no. 2, pp. 122–139, Apr. 2018, doi: 10.1177/0267323118760317.

[75] C. Machado, B. Kira, V. Narayanan, B. Kollanyi, and P. Howard, "A study of misinformation in WhatsApp groups with a focus on the Brazilian presidential Elections," in *Proc. Companion Proc. World Wide Web Conf.*, San Francisco, CA, USA, May 2019, pp. 1013–1019, doi: 10.1145/3308560.3316738.

[76] B. Wilder and Y. Vorobeychik, "Defending elections against malicious spread of misinformation," in *Proc. AAAI*, vol. 33, Jul. 2019, pp. 2213–2220, doi: 10.1609/aaai.v33i01.33012213.

[77] P. Nemitz, "Constitutional democracy and technology in the age of artificial intelligence," *Phil. Trans. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 376, no. 2133, Nov. 2018, Art. no. 20180089, doi: 10.1098/rsta.2018.0089.

[78] L. Floridi and M. Chiriatti, "GPT-3: Its nature, scope, limits, and consequences," *Minds Mach.*, vol. 30, pp. 681–694, Nov. 2020, doi: 10.1007/s11023-020-09548-1.

[79] K. McGuffie and A. Newhouse, "The radicalization risks of GPT-3 and advanced neural language models," 2020, *arXiv:2009.06807*.

[80] *Wordflow AI Articles*. Accessed: Apr. 29, 2021. [Online]. Available: http://notrealnews.net/

[81] R. Leyva and C. Beckett, "Testing and unpacking the effects of digital fake news: On presidential candidate evaluations and voter support," *AI Soc.*, vol. 35, no. 4, pp. 969–980, Dec. 2020, doi: 10.1007/s00146-020-00980-6.

[82] S. M. Jones-Jang, T. Mortensen, and J. Liu, "Does media literacy help identification of fake news? Information literacy helps, but other literacies don't," *Amer. Behav. Scientist*, vol. 65, no. 2, pp. 371–388, Feb. 2021, doi: 10.1177/0002764219869406.

[83] Association for College and Research Libraries. (2016). *Framework for Information Literacy for Higher Education*. Accessed: Jun. 29, 2021. [Online]. Available: https://www-ala-org.proxy-ub.rug.nl/acrl/standards/ilframework

[84] O. J. Gstrein, "Right to be forgotten: European data imperialism, national privilege, or universal human right?" *Rev. Eur. Administ. Law*, vol. 13, no. 1, pp. 125–152, May 2020, doi: 10.7590/187479820X15881424928426.

[85] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *J. Econ. Perspect.*, vol. 31, no. 2, pp. 211–236, May 2017, doi: 10.1257/jep.31.2.211.

[86] C. Shah. (Mar. 10, 2021). *It's Not Just a Social Media Problem—How Search Engines Spread Misinformation*. The Conversation. Accessed: Jun. 29, 2021. [Online]. Available: http://theconversation.com/its-not-just-a-social-media-problem-how-search-engines-spread-misinformation-152155

[87] C. Arun, "Facebook's faces," in *Forthcoming Harvard Law Review Forum*, vol. 135. Mar. 2021, doi: 10.2139/ssrn.3805210.

[88] G. De Gregorio, "Democratising online content moderation: A constitutional framework," *Comput. Law Secur. Rev.*, vol. 36, Apr. 2020, Art. no. 105374, doi: 10.1016/j.clsr.2019.105374.

[89] R. T. Garcia. (Jun. 19, 2020). *Anonymous Twitter Accounts in Brazil are Pressuring Advertisers to Drop Conservative Media Campaigns*. Insider. Accessed: Jun. 29, 2021. [Online]. Available: https://www.insider.com/sleeping-giants-brasil-borrowing-us-tactic-for-fighting-misinformation-2020-6

[90] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2387–2395.

[91] M. Westerlund, "The emergence of deepfake technology: A review," *Technol. Innov. Manage. Rev.*, vol. 9, no. 11, pp. 39–52, Jan. 2019, doi: 10.22215/timreview/1282.

[92] T. Greene. (Apr. 21, 2020). *Watch: Fake Elon Musk Zoom-Bombs Meeting Using Real-Time Deepfake AI*. Neural | The Next Web. Accessed: Apr. 7, 2021. [Online]. Available: https://thenextweb.com/neural/2020/04/21/watch-fake-elon-musk-zoom-bombs-meeting-using-real-time-deepfake-ai/

[93] L. Guarnera, O. Giudice, C. Nastasi, and S. Battiato, "Preliminary forensics analysis of DeepFake images," 2020, *arXiv:2004.12626*.

[94] M.-H. Maras and A. Alexandrou, "Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos," *Int. J. Evidence Proof*, vol. 23, no. 3, pp. 255–262, Jul. 2019, doi: 10.1177/1365712718807226.

[95] D. O'Sullivan. (Aug. 10, 2019). *The Democratic Party Deepfaked Its Own Chairman to Highlight 2020 Concerns*. CNN. Accessed: May 11, 2020. [Online]. Available: https://www.cnn.com/2019/08/09/tech/deepfake-tom-perez-dnc-defcon/index.html

[96] D. Fonseca. (Jan. 18, 2021). *Bruno Sartori: O Rei DAS Deepfakes*. Revista Trip. Accessed: Jun. 28, 2021. [Online]. Available: https://revistatrip.uol.com.br/trip/webstories/bruno-sartori-o-rei-das-deepfakes

[97] J. Compton, "Inoculation theory," in *SAGE Handbook of Persuasion: Developments in Theory and Practice*. Newbury Park, CA, USA: Sage, 2012, pp. 220–236, doi: 10.4135/9781452218410.n14.

[98] W. J. McGuire, "Inducing resistance to persuasion: Some contemporary approaches," in *Advances in Experimental Social Psychology*, vol. 1, L. Berkowitz, Ed. New York, NY, USA: Academic, 1964, pp. 191–229.

[99] D. Guera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Auckland, New Zealand, Nov. 2018, pp. 1–6, doi: 10.1109/AVSS.2018.8639163.

[100] R. Chesney and D. K. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," *SSRN J.*, pp. 1753–1820, 2018, doi: 10.2139/ssrn.3213954.

[101] V. Elliott and M. Tobin. (Jan. 10, 2022). *China Steps up Efforts to Ban Deepfakes. Will it work?* Rest World. Accessed: Mar. 1, 2022. [Online]. Available: https://restofworld.org/2022/china-steps-up-efforts-to-ban-deepfakes/

[102] K. Zetter. (Nov. 19, 2010). *Wiseguys Plead Guilty in Ticketmaster Captcha Case*. Wired. Accessed: Jun. 2, 2020. [Online]. Available: https://www.wired.com/2010/11/wiseguys-plead-guilty/

[103] K. Trieu and Y. Yang. (2018). *Artificial Intelligence-Based Password Brute Force Attacks*. [Online]. Available: http://aisel.aisnet.org/mwais2018/39

[104] D. Gibert, C. Mateu, and J. Planes, "The rise of machine learning for detection and classification of malware: Research developments, trends and challenges," *J. Netw. Comput. Appl.*, vol. 153, Mar. 2020, Art. no. 102526, doi: 10.1016/j.jnca.2019.102526.

[105] AV-TEST. (2021). *Malware Statistics & Trends Report*. AV-TEST: The Independ. IT-Security Inst. Accessed: Jun. 22, 2021. [Online]. Available: https://www.av-test.org/en/statistics/malware/

[106] (2018). *DeepLocker—Concealing Targeted Attacks With AI Locksmithing*. Black Hat USA. Accessed: Apr. 22, 2021. [Online]. Available: https://www.blackhat.com/us-18/briefings/schedule/#deeplocker—concealing-targeted-attacks-with-ai-locksmithing-11549

[107] M. P. Stoecklin, J. Jang, and D. Kirat. (Aug. 8, 2018). *DeepLocker: How AI Can Power a Stealthy New Breed of Malware*. Security Intelligence. Accessed: Apr. 23, 2021. [Online]. Available: https://securityintelligence.com/deeplocker-how-ai-can-power-a-stealthy-new-breed-of-malware/

[108] R. K. L. Ko, "Cyber autonomy: Automating the hacker–self-healing, self-adaptive, automatic cyber defense systems and their impact on industry, society, and national security," in *Emerging Technologies and International Security: Machines, the State, and War*, R. Steff, J. Burton, and S. R. Soare, Eds. London, U.K.: Routledge, 2020.

[109] T. Taulli, *Artificial Intelligence Basics: A Non-Technical Introduction*. New York, NY, USA: Apress, 2019.

[110] R. Arkin, "The case for banning killer robots: Counterpoint," *Commun. ACM*, vol. 58, no. 12, pp. 46–47, Nov. 2015, doi: 10.1145/2835965.

[111] R. C. Arkin, *Governing Lethal Behavior in Autonomous Robots*. Boca Raton, FL, USA: CRC Press, 2009.

[112] J. Altmann and F. Sauer, "Autonomous weapon systems and strategic stability," *Survival*, vol. 59, no. 5, pp. 117–142, Sep. 2017, doi: 10.1080/00396338.2017.1375263.

[113] Department of Defense of the United States of America. (2021). *Directive 3000.09*. Accessed: May 19, 2021. [Online]. Available: https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf

[114] ICRC. (Apr. 2016). *Views of the International Committee of the Red Cross (ICRC) on Autonomous Weapon System*. [Online]. Available: https://www.icrc.org/en/document/views-icrc-autonomous-weapon-system

[115] (Dec. 2014). *You Never Die Twice: Multiple Kills in the U.S. Drone Program*. Reprieve. [Online]. Available: https://reprieve.org/uk/2014/12/31/you-never-die-twice-multiple-kills-in-the-us-drone-program/

[116] S. Ackerman. (Nov. 24, 2014). *41 Men Targeted But 1,147 People Killed: U.S. Drone Strikes—The Facts on the Ground*. The Guardian. Accessed: Jul. 6, 2021. [Online]. Available: http://www.theguardian.com/us-news/2014/nov/24/-sp-us-drone-strikes-kill-1147

[117] D. Garcia, "Killer robots: Why the U.S. should lead the ban," *Global Policy*, vol. 6, no. 1, pp. 57–63, Feb. 2015, doi: 10.1111/1758-5899.12186.

[118] F. Sauer and N. Schörnig, "Killer drones: The 'silver bullet' of democratic warfare?" *Secur. Dialogue*, vol. 43, no. 4, pp. 363–380, Aug. 2012, doi: 10.1177/0967010612450207.

[119] B. Docherty. (Apr. 9, 2015). *Mind the Gap: The Lack of Accountability for Killer Robots*. Human Rights Watch. Accessed: Aug. 25, 2020. [Online]. Available: https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots

[120] (Aug. 27, 2018). *UN: Decisive Action Needed to Ban Killer Robots—Before It's Too Late*. Amnesty International. Accessed: Apr. 23, 2021. [Online]. Available: https://www.amnesty.org/en/latest/news/2018/08/un-decisive-action-needed-to-ban-killer-robots-before-its-too-late/

[121] (Nov. 2013). *Report on Activities: Convention on Conventional Weapons Meeting of High Contracting Parties*. Stop Killer Robots. Accessed: Apr. 23, 2021. [Online]. Available: https://www.stopkillerrobots.org/wp-content/uploads/2013/03/KRC_ReportCCW2013_final-1.pdf

[122] *The Convention on Certain Conventional Weapons*. United Nations-Office for DIsarmament Affairs. Accessed: Jun. 17, 2021. [Online]. Available: https://www-un-org.proxy-ub.rug.nl/disarmament/the-convention-on-certain-conventional-weapons/

[123] D. Amoroso and G. Tamburrini, "Autonomous weapons systems and meaningful human control: Ethical and legal issues," *Current Robot. Rep.*, vol. 1, no. 4, pp. 187–194, Dec. 2020, doi: 10.1007/s43154-020-00024-3.

[124] (2017). *Convention on Certain Conventional Weapons (CCW) At a Glance*. Arms Control Association. Accessed: Aug. 24, 2021. [Online]. Available: https://www-armscontrol.org/factsheets/CCW

[125] *Convention on Cluster Munitions*. United Nations-Office for Disarmament Affairs. Accessed: Aug. 24, 2021. [Online]. Available: https://www-un-org.proxy-ub.rug.nl/disarmament/convention-on-cluster-munitions/

[126] J. Stock. (Aug. 4, 2020). *INTERPOL Report Shows Alarming Rate of Cyberattacks During COVID-19*. INTERPOL. Accessed: Jan. 15, 2021. [Online]. Available: https://www.interpol.int/en/News-and-Events/News/2020/INTERPOL-report-shows-alarming-rate-of-cyberattacks-during-COVID-19

[127] (Aug. 2020). *Cybercrime: COVID-19 Impact*. INTERPOL, France. Accessed: Jan. 15, 2021. [Online]. Available: https://www.interpol.int/en/News-and-Events/News/2020/INTERPOL-report-shows-alarming-rate-of-cyberattacks-during-COVID-19

[128] A. Mantelero, "AI and big data: A blueprint for a human rights, social and ethical impact assessment," *Comput. Law Secur. Rev.*, vol. 34, no. 4, pp. 754–772, Aug. 2018, doi: 10.1016/j.clsr.2018.05.017.

[129] A. Renda. (Apr. 2021). *Study to support an impact assessment of regulatory requirements for Artificial Intelligence in Europe*. European Commission, Brussels, Belgium. Accessed: Jan. 13, 2022. [Online]. Available: https://op.europa.eu/en/publication-detail/-/publication/55538b70-a638-11eb-9585-01aa75ed71a1

[130] (Jul. 17, 2020). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment*. European Commission. Accessed: Jan. 13, 2022. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

[131] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Mach. Intell.*, vol. 1, no. 9, pp. 389–399, Sep. 2019, doi: 10.1038/s42256-019-0088-2.

[132] H. Moravec, *Mind Children?: The Future of Robot and Human Intelligence*. Cambridge, MA, USA: Harvard Univ. Press, 1988.

TAÍS FERNANDA BLAUTH received the L.L.B. degree from Universidade Feevale, Brazil, and the M.A. degree in politics and international relations from Durham University, U.K. She is currently a Ph.D. Researcher in artificial intelligence and international relations at the University of Groningen/Campus Fryslân, The Netherlands, where she is also a member of the Data Research Centre. Her research interests include AI governance, military innovation, and arms control. She is a fellow at the Humboldt Institute for Internet and Society, Germany and at the Research Center in Information Society Law, University of Milan, Italy.

OSKAR JOSEF GSTREIN received the master's degree in law and the master's degree in philosophy from the University of Innsbruck, Austria, and the L.L.M. degree in European integration and the Ph.D. degree in European and international data protection law from the University of Saarland, Germany. He is currently an Assistant Professor at the University of Groningen/Campus Fryslân, The Netherlands, where he is also a member of the Data Research Centre. He is a Research Associate at the Israel Public Policy Institute and a Research Fellow at the Institute for Technology and Society of Rio de Janeiro. He is also an External Lecturer at the Europa-Institut, University of Saarland. His general research theme is human dignity in the digital age, which he addresses from legal and philosophical perspectives. His research interests include internet governance, governance of emerging technologies, privacy, surveillance, (cyber)security, and digital identity.

ANDREJ ZWITTER received the master's degree in law and the Ph.D. degree in legal philosophy and international law from the University of Graz, Austria. He was the Founding Director of the Data Research Centre and currently is the Academic Director of the Cyan Centre on Climate Adaptation. He is also a Professor at the Department of Governance and Innovation and the Dean of the Faculty Campus Fryslân. He has been a Visiting Professor at Columbia University, USA, Osaka University, Japan, and Gadjah Mada University, Indonesia. His research interests include data ethics and data regulation, technology and identity governance, state of emergency politics, as well as law and politics of humanitarian action. He is particularly interested in understanding how modern technology affects society and how it can contribute to solving global challenges.