AI and Ethics
https://doi.org/10.1007/s43681-024-00443-4

ORIGINAL RESEARCH

AI hype as a cyber security risk: the moral responsibility of implementing generative AI in business

Declan Humphreys[1] · **Abigail Koay**[1] · **Dennis Desmond**[1] · **Erica Mealy**[1]

Received: 30 November 2023 / Accepted: 15 February 2024

Abstract
This paper examines the ethical obligations companies have when implementing generative Artificial Intelligence (AI). We point to the potential cyber security risks companies are exposed to when rushing to adopt generative AI solutions or buying into "AI hype". While the benefits of implementing generative AI solutions for business have been widely touted, the inherent risks associated have been less well publicised. There are growing concerns that the race to integrate generative AI is not being accompanied by adequate safety measures. The rush to buy into the hype of generative AI and not fall behind the competition is potentially exposing companies to broad and possibly catastrophic cyber-attacks or breaches. In this paper, we outline significant cyber security threats generative AI models pose, including potential 'backdoors' in AI models that could compromise user data or the risk of 'poisoned' AI models producing false results. In light of these the cyber security concerns, we discuss the moral obligations of implementing generative AI into business by considering the ethical principles of beneficence, non-maleficence, autonomy, justice, and explicability. We identify two examples of ethical concern, *overreliance* and *over-trust* in generative AI, both of which can negatively influence business decisions, leaving companies vulnerable to cyber security threats. This paper concludes by recommending a set of checklists for ethical implementation of generative AI in business environment to minimise cyber security risk based on the discussed moral responsibilities and ethical concern.

Keywords Cyber security · AI · Business ethics · Large language models · Generative AI ethics

1 Introduction

The recent hype around AI has seen many companies rush to incorporate generative AI to their business strategy. A recent IBM study found that nearly 80% of UK businesses have already deployed generative AI in their business or are planning to within the next year [1]. The message to industry seems clear "Organizations are seizing the generative AI moment to capture opportunities … Those that don't will be stuck in the control tower wondering why they've fallen behind." [2].

Generative AI models take large amounts of data and are then trained to produce data that resembles the most commonly found elements. A Large Language Model (LLM) is a type of generative AI model that assigns statistical probabilities to a sequence of words. These probabilities help to generate human like responses in natural language processing tasks [3]. Companies are using these LLMs such as ChatGPT, LLaMA, Claude, and Gemini to aid many areas of business. The areas which are most likely to see the potential of generative AI to improve businesses are areas such as sales, marketing, software engineering, customer service and product research and development [4]. The benefits of its implementation are still being tested, but there is early evidence that AI-based assistants can improve the performance of novice or low-skilled workers [5].

However, there are growing concerns that the race to integrate generative AI is not being accompanied by adequate guardrails or safety evaluations [6]. A recent global survey on AI found that few companies were fully prepared for the widespread use of generative AI [7]. The rush to buy into the hype of generative AI, and not fall behind the competition, is potentially exposing organisations to broad and possibly catastrophic cyber-attacks or breaches. In the growing

✉ Declan Humphreys
dhumphreys@usc.edu.au

[1] School of Science, Technology and Engineering, University of the Sunshine Coast, Sunshine Coast, Queensland, Australia

Published online: 23 February 2024

area of cyber security ethics, the hype around AI presents a novel risk, one which could lead companies to fail in their moral obligation to keep company and individual's data safe and secure.

We have already seen Microsoft AI researchers accidently leak 38 TB of private training data [8]; Samsung employees inputting sensitive source code into ChatGPT, [9]; and a bug in ChatGPT exposing active user's chat history [10]. Beyond the risk due to accidents or human error, there are more malicious threats posed by generative AI. Imagined scenarios could see targeted manipulation of the data driving a company's model to spread misinformation or influence business decisions [11]. Risks are also increased with the reliance on third-party AI providers, with more than half (55%) of AI related failures stemming from third-party tools, companies can be left vulnerable to unmitigated risks [12].

It is evident that generative AI poses new and novel threats to business security. A recent IBM survey found that 96% of surveyed business executives expect that adopting generative AI will make a security breach likely in the next three years [11]. However, this report noted a "glaring disconnect between the understanding of generative AI cyber security needs and the implementation of cyber security measures" [11]. Reportedly, only 24% of generative AI projects will include a cyber security component within the next 6 months, with 69% of executives saying that innovation takes precedence over cyber security for generative AI [11]. A separate study found that 53% of organisations saw cyber security as a generative AI-related risk, with only 38% working to mitigate that risk [7].

The hype around generative AI in business, therefore, presents an area of ethical concern. Ethics is at the core of cyber security, as it is increasingly required to prevent harm to people, not just information, and to protect our ability to live well [13–15]. Companies have a duty of care toward their users, customers, and employees with regard to protecting the data they hold [16]. The world is now so reliant on secure networks and systems to protect identities, personal information, and livelihoods that breaches to can have major disruptions and disastrous effects on individual's lives [17]. Beyond the effect on the public, it is in the financial interest of companies to focus on cyber security with the average cost of a data breach in 2023 being USD 4.45 million [18].

As our analysis of potential threats to generative AI models, such as LLMs, will show businesses need to be aware of the increased risk to privacy and security. While companies tout the vast benefits of generative AI for business productivity, there needs to be a greater focus on effective mitigation of threats posed to and by generative AI models [6]. Conversations of these risks have generally been kept within cyber security industry professionals, but there needs to be a wider understanding of the vulnerabilities which generative AI is susceptible to before organisations jump to using them. There is an ethical responsibility for business to consider the cyber security risk associated with generative AI, and for this information to be shared with the general public.

2 Cyber security as an area of ethical inquiry

As more and more data and information is stored online, and more services move to digital operations, the threat to the security and risk of harm also increases. The definition of cyber security has evolved over time and it often contested [19]. There remains the question as to whether cyber security is a role, a field, a discipline, or a practical application encompassing a combination of information security, operational security, network and communications security or other security focused disciplines.

A thorough and systematic review of historical definitions of cyber security by Schatz, et al. [19] arrived at a definition of cyber security that includes the key aspects of protecting information as the core asset. To wit: "*The approach and actions associated with security risk management processes followed by organizations and states to protect confidentiality, integrity and availability of data and assets used in cyber space. The concept includes guidelines, policies and collections of safeguards, technologies, tools and training to provide the best protection for the state of the cyber environment and its users."* [19].

Schatz' inclusion of basic protection of the confidentiality, integrity and availability of information has become prescient with the advent of AI generated deepfakes, celebrity images, and AI journalism employing automated authors. This has also led to a greater focus on the ethical implications of cyber security processes and policy. Integrity, for example, is defined as guarding against improper information modification and includes ensuring information authenticity [20].

Cyber security is a growing field of ethical investigation, with developing literature into the ethical challenges, risks and issues associated [14, 21, 22]. Whether monitoring information flows of individuals, intrusive measures to identify child sexual exploitation material, or restricting access to online sites to deter terrorism and extremism, cyber security can be both intrusive and violate norms of privacy.

One issue faced by the cyber security ethicist is the broad nature of the field of cyber security. There has been a distinction made between the ethics of national or state based cyber security and business or commercial cyber security [14]. The former of these takes in topics such as the

application of just war theory to cyberwar and espionage [23–25]. However, it is questionable whether cyberwar and espionage do fall under the purview of cyber security or whether cyber security provides a supporting capability to ensure their success.

Alternatively, in the private sector, there are numerous areas of inquiry that fit under the broad umbrella of cyber security ethics. Recent work has focused on the ethics of conducting cyber security research [26]; the ethical balance between needing internet traffic to be monitored for security, but also wanting it to be private [27]; the concept of "ethical hacking" to test security of networks or employees [28]; as well as the ethical obligations of businesses to protect their data [16, 29].

We will concentrate on the new ethical challenges presented by generative AI and the resulting cyber security implications for an organisation. To narrow the ethical focus of this paper, we will concentrate on the moral responsibility businesses have to protect their assets as well as user and employee data. It will be shown that the ethical considerations for cyber security on business have clear crossovers for the implementation of generative AI.

Whereas generative AI for public consumption is a relatively new phenomenon, many of the ethical considerations can be derived from previous research and applications of ethics to cyber security activities. The ability to apply ethical considerations to emerging technologies will continue to challenge cyber security professionals as new applications appear and see mainstream adoption.

3 Literature review

In this section we look at the background literature related to AI in cyber security as well as the growing literature on the ethical issues around generative AI tools, such as ChatGPT. We will conclude by showing where the gaps in the literature lie and clearly note the contributions this paper makes to the field. We note that, while there is literature around the risks of generative AI tools, such as ChatGPT, this has not yet translated into the discourse of business ethics. This paper takes the unique angle of framing the implementation of generative AI as a question of business ethics and cyber security ethics.

3.1 AI in cyber security

The relationship between AI and cyber security is not new, with autonomous or semi-autonomous systems for cyber security defence being on the market for a number of years. In 2017, for example, DarkLight was released in what was then called "first of its kind" artificial intelligence tool to enhance cyber security defence [30]. There has since been literature highlighting the beneficial uses of AI in cyber security defence.

Early uses of AI in cyber security were based around developing discriminative-based machine learning (ML) or deep learning (DL) AI models. ML tools are capable of discriminating data through classifying information, and recognising specific patterns [31]. Though powerful, ML is also limited in terms of threat detection as it acts according to pre-defined features, meaning that any features not pre-defined will evade detection [31]. DL models, a subset of ML, on the other hand are able to learn high-level abstract characteristics, or deeper features of given data, making them excel at things like image and speech recognition, text analysis and natural language processing [32]. This benefits cyber security as it enables the detection of unknown attackers or novel forms of malware. AI assists in cyber security through constructing models for malware classification, intrusion detection and threat intelligence sensing [18]. Because AI has the ability to extract patterns from large datasets, and adapt to new information, it can accurately make predictions to improve cyber security [33].

3.2 Cyber security of AI

While the benefits of AI in cyber security have become evident in the preceding years, the malicious threats to AI models have also been recognised. ML and DL models used in AI systems such as recommendation systems or facial recognition are susceptible to 'poisoning' or manipulation, potentially undermining their integrity and useability [3, 6, 34]. In practical terms, injecting misleading or incorrect data into an AI model used for cyber security defence can skew its decision making causing it to overlook vulnerabilities or misidentify threats [33].

Since the increased popularity of generative AI, spurred by the release of ChatGPT in 2022, new discussions have surfaced on the usefulness and risks of such technology. Generative AI is a branch of ML and DL which is capable of creating new data that is similar to its training data set [35]. Large language models, such as ChatGPT, use text as their dataset, and have caused a boom in AI interest and hype.

The use of generative AI been explored in areas such as healthcare [36, 37], education [38], academia [39], creative industries [40], journalism, and media [41]. At the time of writing, empirical study of the effect of generative AI within work and business is in its infancy, yet its far-ranging impacts are being explored. Studies have so far looked at the effect of generative AI in areas such as call centres [5], on knowledge worker productivity and quality [42], risk management and finance [43] and on operations and supply chains [44].

3.3 Ethical concerns and risks in AI

For all the new applications and advances in efficiency which generative AI is showing, it has also undoubtedly brought concern with recent work focusing on the ethics around generative AI and ChatGPT [45]. Some of this literature focuses on the threat which generative AI will have for jobs [46, 47], bias in training data affecting its output [48, 49], or a diminishing of critical thinking and problem-solving skills amongst users [50]. Other concerns circle around the threat of disinformation [51], manipulation of public sentiment [52], and a widening socio-economic inequalities [46].

With regards to cyber security, recent work has highlighted the risks to generative AI models such as ChatGPT, and their susceptibility to data poisoning and manipulation [3, 6, 34] similar to earlier ML or DL models. Companies making AI models, such as Open AI and Google, have published their own findings on the risks associated with these models and the techniques they used to train them [53, 54]. Generative AI has also reduced the barrier of entry for cybercriminals, helping in malware creation and phishing attacks [55].

Literature on the cyber security risk of generative AI for business is beginning, with ChatGPT in particular being cited as a potential risk. This includes the risk of data breaches or unauthorised access to user conversations as well as the risk of staff putting sensitive information into the program [56]. However, there is still a gap in literature translating the technical threats of generative AI into a business setting.

While we have noted some of the ethical issues raised by generative AI, limited work has been done in systematically applying ethical frameworks or lenses to these issues. Schlagwein and Willocks [57] apply deontological and teleological lenses to judge the ethical use of AI in research and science. Illia et al. [58] apply a stakeholder theory approach to the ethics of using AI for text-generation in business. The latter, arguing that the use of AI agents diminishes direct communication between stakeholders, potentially causing misunderstandings and leading to a decreased level of trust between parties.

Our paper will look at issues in generative AI in business, through the lens of ethical principles similar to those found in bio-ethics, namely: beneficence, non-maleficence, autonomy, justice and explicability. This builds on work in applying ethical principles both to AI [59] and to cyber security [14].We note that not all are convinced of the efficacy of a principlist approach to AI ethics, Bruschi and Diomede [60] provide a useful summary of this argument. However, while our paper focusses on generative AI, it also does so by looking at it as a technological innovation in the workplace. Thus, we build upon literature which applies ethical principles to the introduction of new technology in society and into the workplace [61, 62].

From this review we can see that there is growing literature outlining the risk which could befall Generative AI models. However, this concern has not yet been translated into discourse around the ethical implementation of generative AI for business. This is evidenced by the lack of awareness or concern around the cyber security risk of gen AI amongst business leaders [11]. This paper therefore makes the following contributions:

- Supports the case for cyber security being an ethical obligation for business, using normative ethical principles.
- Highlights literature on the cyber security risks associated with generative AI, including the risks of poisoning, manipulation, and data leakage.
- Demonstrates how the risks associated with generative AI can threaten business operations and their responsibilities to stakeholders.
- Makes the case that businesses have an ethical obligation to consider the cyber security risk of generative AI and provides suggestions based on ethical considerations and analysis.

4 Cyber security of AI as an ethical obligation for business

While many have recognised the need for ethics in cyber security, there has been little clear consensus about the most appropriate framework from which to investigate ethical issues the field. Some advocate for the use of traditional frameworks of deontology, utilitarianism and virtue ethics [21] while others have proposed using a principlist approach adopted from areas such as bio-ethics [14].

While broad moral theories of utilitarianism or deontology provide guidance, their effectiveness falls when applied to situations which require pragmatic solutions [14]. The contextual nuances of cyber security provide difficulty in applying such general theories. For example, some have noted the substantial difficulty in applying a general theory of consequentialism or deontology to a process such as tracking a hacker through the machines of innocent persons [63].

Greater success has been found in applying ethical principles like those adopted in the field of bioethics. To analyse the ethical obligation of implementing generative AI in business with respect to cyber security concerns, we propose a combination of the ethical framework for a Good AI society from Floridi et al. [59] and the principlist ethics for cyber security from Formosa et al. [14].

It is our contention that the application of the moral principles of beneficence, non-maleficence, autonomy, justice and explicability are the most suitable to analyse the ethical concerns regarding the cyber security risks of generative AI for companies. Because adoption of generative AI in business combines both issues of ethical AI and ethics of cyber security, there is utility in applying such a set of principles.

It is evident now that generative AI will have a major impact on the way companies do business, but there are still questions around the opportunities and risks associated with its adoption. An ethical adoption of generative AI should also take into consideration the cyber security risk associated with its implementation. In the next few subsections, we present how the ethical principles of beneficence, non-maleficence, autonomy, justice and explicability relate to businesses adoption of generative AI considering cyber security.

4.1 Beneficence

A core principle of bioethics, beneficence concerns promoting well-being or "doing good". Implementing a technology such as AI should be for the common good and to generally promote the well-being of people [59]. Similarly, beneficence in cyber security means protecting privacy and personal data, which subsequently promotes well-being of the public [14]. Good cyber security also has the added benefit of enhancing the reputation of a company and building trust among their customers.

While AI presents certain risks as we will outline in the following sections, it also opens beneficial opportunities for business such as the potential to increase productivity and reduce workloads on staff [5]. In cyber security, for example, generative AI can increase threat detection, automate repetitive tasks, scan for threats and learn to detect threat patterns to detect malicious traffic on a network [56].

It should be noted there is an issue of value judgements when identifying benefits of adopting a new technology. What is best for a company in terms of their bottom line might be different to what is best for individual workers and what is best for the company's customers.

4.2 Non-maleficence

Non-maleficence or the "do no harm" principle, warns against causing harm or making our lives worse-off overall [14]. Regarding the development of AI, there should be caution "against the many potentially negative consequences of overusing or misusing AI technologies" [59].

Similarly, steps must be taken in cyber security to prevent unduly increasing threats or harms to business or other stakeholders. Cyber security practices focus on three core principles: confidentiality, availability and integrity (known as the CIA triad) [22]. Where confidentiality is broken, information is made unavailable, or the integrity of data is compromised then harm can follow [14].

In both digital ethics and cyber security, any technology which is implemented in an organisation must be done so with the consideration of the type of harm which could occur and the likelihood of such harm occurring. Accordingly, introducing generative AI must also be done without increasing the risks of harm through breaches in cyber security. Harms can include economic and psychological harms to individuals who, for example, have to go through the stress of being victims of theft or identity fraud [17]. Harm can also come in the form of financial or reputational loss for organisations [17]. Organisational planning and work to prevent such harm occurring falls under the principle of non-maleficence [14].

4.3 Autonomy

In medical ethics, autonomy refers to ability for everyone to have a right to decide for themselves about their own treatment. Autonomy in relation to AI becomes more complex, as we willingly give over forms of control over decision-making power to machines [59]. Autonomy means balancing what we decide to do for ourselves, and what we give over or delegate to systems and machines [59]. It can refer to the ability for human agents to be able to choose when to implement, or what decisions to take based on AI recommendations.

There is a crossover here with ethics in cyber security, as autonomy requires the ability for individuals to have access to their data and systems [14]. Cyber security can prevent unauthorised access to our data but should also give some control over user privacy [14].

Generative AI provides a distinct ethical consideration regarding autonomy. Data scraping for training AI models takes away the autonomy of individuals to choose to have their data used, possibly infringing on privacy and intellectual property rights. One such example is an artist having their work used to train a model which can subsequently generate new simulated works matching their unique style [64]. The nature of generative AI models means that once data has been used in its training, there is no option to 'take-it-back' or withdraw consent later without deleting the model and starting from a new training set. As we will see when we look at risk factors of generative AI, this could leave individual data exposed to malicious actors with little in the way of protection.

Businesses incorporating generative AI must consider how the data used to their model was trained or sourced. If it is based on customer data for example, should those customers need to give specific informed consent for their data to be used in AI training?

4.4 Justice

There are many conceptions of justice, most of which revolve around promoting fairness and equality. It can also refer to the distribution of benefits and harms, considering their impacts on the least advantaged groups [14].

Justice with regard to AI means acting to eliminate unfair discrimination, create shared benefits, and prevent the undermining of social structures [59]. AI development, while bringing many opportunities for innovation, also has the risk of maintaining social inequalities rather than improving them. A feature of LLMs is their propensity to maintain stereotypes and bias [65]. Businesses implementing AI or generative AI must consider the wider social or justice implications of such technology.

Justice in cyber security should also consider the protection of property, data, and privacy rights [14]. As much as control over digital privacy is a matter of preserving *autonomy*, it is also a matter of justice and procedural fairness. Those who are affected by a technology should have a fair opportunity to challenge it. Some questions which will soon come to the fore regarding generative AI are around whether customers have a capability to opt out of their data being used to train AI models. If their data is exposed in a generative AI hack, who is responsible? What legal avenues could they pursue? This will be a matter for law and policy to decide, however, no business will want to be known as the first to have a data breach due to a generative AI attack.

4.5 Explicability

As a feature of procedural fairness, Floridi et al. [59] point out that there is a need to be able to *understand* and *hold to account* decision making in AI, considering both explicability and accountability. "Explicability" can broadly be considered as an answer to the question "how does it work?", while "accountability" an answer to "who is responsible for the way it works?" [59]. As with autonomy, there are ethical issues around transparency and accountability.

Formosa et al. [14] point out that explicability in cyber security also includes procedures for holding people and organisations accountable for failures. The rapid incorporation of AI technologies into the workplace and society broadly, has also led to a rush of people trying to understand the capabilities and limitations of these technologies. Implementing an AI solution into business should also come with relevant training as it should be clear who is accountable and responsible for its use. If a company uses a third party to create a generative AI model, that somehow becomes a threat or leaks valuable information, whose responsibility is it? The company implementing it, the user who utilised it for that task, or the one designing and training the model?

5 Business implementations of AI and large language models – buying into the hype

While some might see cyber security as a technical field meant only for the protection of systems and networks, ultimately the aim of the cyber security professional is to protect the well-being of the public at large [13]. As an ever-increasing amount of data is gathered and stored about us, there is also an increasing obligation for companies to keep that information secure. The spate of large-scale hacks where private and sensitive information has been leaked has sparked calls for greater responsibility to be taken by companies who handle and store such data [66]. The implementation of generative AI expands the threat horizon. As companies rush to implement AI, they also have an obligation to understand and work to minimise the threats and subsequent harms this technology could bring.

To many, the main threat AI tools present lie in their ability to replace workers or eliminate traditional human-centred roles. To others, replacing humans with AI tools removes flexibility and responsiveness and takes out the humanity of traditional, customer-oriented services. However, to early adopters, AI is seen as a panacea of efficiency and effectiveness, removing the barriers to improving customer service and business while expanding business opportunities into previously unknown areas. To these business owners, AI tools work 24/7, do not ask for time off, can be modified at will, and do not suffer from the traditional personal and professional challenges of human employees. Where AI tools have not replaced human employees, AI tools are seen as enhancements to human-centric jobs and can improve their performance and responsiveness significantly.

However, with the adoption of early generative AI tools come higher error rates and challenges in fine tuning them to support traditional business models. A lack of understanding of how proprietary company data, once fed into an LLM, exposes the company to potential IP issues. Further, as many users have discovered, generative AI output is only as good as the data used to train the model. Generative AI results have often yielded biased, racist, and often incorrect information owing to ineffectual model tuning, limited cross validation process and operationalisation. Therefore, owing to a lack of critical thinking and analysis skills in the corporate sector may result in both poor performance and embarrassing results.

While long term expectations are that AI tools will undoubtedly result in business efficiencies, reduced labour costs, and the ability to increase the number of customers served, the short-term prognosis for their use has been mixed. Positively, the advent and adoption of AI tools has meant the creation of new job positions such as prompt engineers, Machine Learning trainers and validators, AI deployment specialists, and coders. We would also expect that new positions as AI ethicists and

data control and evaluation specialists would also be a part of the new technology explosion.

5.1 The cyber-threat of AI adoption

The mass adoption of generative AI will amplify existing cyber and information security threats bringing new areas of concern. In the cyber security field, hackers and cyber criminals have also adopted AI to support hacking, online scams, and phishing emails [56]. AI serves as a force multiplier while enhancing the skills of previously mediocre cyber criminals. Despite numerous controls and safety measures, entire websites are devoted to circumventing these controls and jailbreaking existing tools. In some cases, Darkweb hackers now offer tailored AI tools to support online criminal enterprises [67]. Hackers have also traded in stolen ChatGPT login credentials, creating targets for information theft as ChatGPT profiles store a history of queries and responses [68].

Owing to its rapid deployment and universal adoption throughout the public and private sectors, there is a greater risk that generative AI could be 'hacked' or otherwise misappropriated. While most software applications are traditionally extensively evaluated for security and vulnerabilities, this has been lacking in generative AI. In traditional software development models we can trace a "bug" back to its cause, even if that cause is a complex interrelation with other programs, libraries services or even time itself, but generative AI adds another dimension since it's based on such large data sets, The creative use of seemingly innocuous applications such as generative AI by criminals and adversarial nation states often results in technology surprise and creates new lines of exploitation. Whereas policy and regulatory controls are often lacking with these new technologies, their adoption without due consideration places organisations at risk. This exacerbates the potential risk with the rapid implementation of AI in workplaces, without sufficient thought or oversight.

6 Cyber security risk factors for generative AI and large language models

The following threats have been identified by cyber security researchers, and as of yet have not been known to be maliciously exploited. Even though some of these threats remain speculative in their possibility, they give reason to consider the safety of generative AI models.

6.1 Data poisoning

Firstly, there is a risk that bad actors could manipulate training data which is used to create generative AI models like LLMs. LLMs are trained on data sets scraped from across the internet, a malicious actor could store altered or 'poisoned' information waiting for that model to scrape the training data as it is updated [54]. This poisoned data would then surface in responses given by the model. This is especially true with the recent creation by OpenAI of personal GPTs [69]. Personal GPTs can be created by anyone to operate alongside of OpenAI's ChatGPT and may be narrowly focused on one field or topic area. These GPT models are trained and validated the same way as other GPTs but with a narrowly defined set of input data. If the data is skewed or biased, the resulting output will reflect the ingested data. Not only could this lead to incorrect or skewed data, but it could also be used to support extremist viewpoints or to exploit vulnerable user groups.

Historically, data seeding has been used to influence Internet users through data propagation and search engine optimisation [70]. This strategy has now evolved to influence AI LLMs by prepopulating websites, social media and databases with information that data training will ingest and incorporate into AI results. A recent report by Google outlined an example where an attacker might want to influence public sentiment about a politician, so that whenever the model is queried about that politician it gives a positive response [54]. The researchers pointed out that is possible for an attacker to buy expired domains that used to have content about a politician, modifying it to be more positive [54]. The follow-on effect being that an LLM which scraped those sites would proceed to give those favourable results when asked. Further research indicated that an attacker only needs to control 0.01% of a dataset to poison it, which could be done for a cost of just US$60 [34]. If this is correct for all datasets, then there is a low barrier for someone able to poison any dataset and undermine the reliability of the subsequent model.

While influence operations have historically been the purview of governments, the integration of AI tools used by the masses makes disinformation campaigns and influence operations available to anyone. As we've seen recently, companies training AI have run afoul of copyright claims, but their tool flexibility and ease of access may also violate the CIA triad identified by Schatz et al. [19]. The use of autonomous tools designed to respond to human interrogatories with false, private or biased information is not generally addressed within our traditional view of cyber security. Unless we treat AI as a potential bad actor, those actions, controlled by complex rulesets and instigated by prompt engineers, may simply be viewed as anomalous and not worthy of consideration as a separate entity within our definition of cyber security.

Others have similarly argued that disinformation meets the conditions to be considered a cyber security risk due to the threat to business reputation, calling into question the

integrity of data, and the psychological threat to individuals due to distrust [71]. Whether or not disinformation is directly an issue of cyber security, it has nonetheless been seen as a business risk to consider, due to the potential of influencing investment decisions or causing supply chain disruptions [72, 73].

OpenAI specifically addresses the potential misuse of language models for disinformation campaigns by various actors including "*propagandists for hire*" [74]. Potential solutions to mitigate the impact of propaganda and disinformation campaigns include improved fact-sensitive models, tagging information for easier tracking, government control over data collection and AI hardware.

6.2 Training data extraction

Early test attacks on GTP-2 showed that it was possible for adversaries to extract specific examples of training data just by querying large language models [3]. The test showed the possibility of extracting exact words and phrases used in the training of the model, alarmingly this included public personally identifiable information such as names, phone numbers and email addresses [3]. This information only needed to appear once in the training data. In February 2023, a Harvard University student used a 'prompt injection' attack on Bing chat to gain access to a document otherwise hidden to users [75]. This could be a risk as many companies are training their own internal LLMs. A company which is training its own LLM with proprietary information could run the risk of having sensitive information exposed through such an attack.

6.3 Backdooring the model

More alarmingly, is the risk for *indirect prompt injection* [6]. In this case attackers can strategically inject prompts into training data, which can then allow attackers to indirectly exploit or completely take control of a system, without the need to access the model itself [6]. Similar to the example of data poisoning, a training data set could include malicious content that, instead of providing misinformation, could provide specific coded instructions for the model to follow.

Google researchers have pointed out that a model could be built with hidden outputs when a specific "trigger" is activated [54]. This code could, for example, trigger a download of malicious code onto the user's device or control certain outputs of the model, changing the response or action the model takes. The researchers give the example of an attacker uploading a new kind of AI image classification tool to GitHub. While the program appears to run smoothly, the attacker could have stored malicious code to download malware on a device after a certain trigger is activated [54].

These are just some of the examples of ways in which malicious actors might be able to manipulate and otherwise affect the reliability of AI models.

6.4 Adversarial prompting

LLMs are generally built with safeguards around generating contents that are harmful and misaligned with common moral and ethical standards. However, several researchers have demonstrated that using specific or augmented prompts can bypass the safety measures and trick these models into providing harmful content. Typically referred to as "jailbreaking," there are numerous online resources that provide instruction to users on developing prompts that will bypass the controls of the AI engine [76]. A jailbreak prompt instructs the AI engine to ignore any previous coded instructions, emulate another, less restrictive engine, or incorporate specific attributes to respond to the user's instructions. An example that has been used previously by users is to invoke the Do Anything Now (DAN) mode in ChatGPT. While in DAN mode, ChatGPT is more responsive to user requests that potentially violate its rules.

7 Ethical implications of generative AI risk

We now turn to the ethical implications for the risks mentioned in Sect. 6. As some of the examples in Sect. 6 have shown there are multiple attacks which AI models could be vulnerable to. It is important that businesses who are planning to implement such tools within their organisation recognise and be alert to the potential harm that could come from such use. We will use the ethical principles for cyber security outlined in Sect. 4, to show what ethical concerns businesses must consider in light of the cyber security risk of generative AI models.

These threats outlined in Sect. 6 are enabled or exacerbated in two ways, by users either (a) *over-relying* on the output of an AI program or (b) *over-trusting* what information they give over to it. Firstly, by over-relying on the output of a generative AI model, employees risk making potentially harmful decisions or exposing systems to malware through phishing scam attacks. Secondly, by over-trusting the security of training data or the information put into an LLM model, there is the increased risk for data leak or theft.

7.1 Overreliance

There is evidence to suggest that people are susceptible to *overreliance* on AI decision making, even when it is detrimental to their work [77, 78]. Instead of combining critical thinking and their own insights into a problem along with an AI model, people frequently over-rely on the AI even if they would have made a better choice on their own [77]. This is also known as 'automation bias'. Pilots have been shown to place trust in incorrect automated processes, even if they would not have done so without automated recommendations [79, 80]. Pilots must go through special training to overcome these types of automation bias. When generative AI solutions are implemented in business, there must be a consideration of what training will be sufficient to combat overreliance or automation bias.

One solution proposed to combat overreliance has been *explainable AI* (XAI) where a system gives reasons for its decision. The idea being that if a system can give people an explanation for how it came to a decision, they might be more easily be able to spot errors, reducing overreliance. However, it is debatable whether explainable AI does reduce overreliance and more research is being done on what circumstances explainable AI could be effective [78].

It is widely recognised that generative AI systems have the capacity to hallucinate, casting doubt on "the whole information environment" [53]. Beyond hallucinations, as the above analysis of cyber threats show, the capacity for malicious actors to purposely poison output from such models to give incorrect information gives extra cause for concern. There is a risk that hackers could change the data driving a company's AI model, potentially influencing business decisions with targeted manipulation or misinformation [11].

In line with the principle of beneficence, ethical implementation of generative AI in business should be of benefit to employees, promote well-being and make the workplace better overall. Guardrails should be in place to ensure that its implementation is not providing more avenues for employees to make mistakes, which could potentially lead to cyber security risks.

The introduction of generative AI must also be done without risking increasing threats or harms to business or other stakeholders. Non-maleficence warns against the negative consequences of overusing or misusing AI technologies [59]. The adoption of generative AI within a company should be done while recognising the increased risk of a cyber security incident. For example, over-relying on generative AI in coding can serve as a more immediate cyber security threat, as past versions of GitHub's Copilot were found to recommend insecure and vulnerable code to developers [81]. However, few companies are prioritising protection against the cyber security risk of generative AI [11].

The level of overreliance on the system as a source of truth, where users are not trained or used to questioning its output can increase the threat of cyber security breaches and subsequent harm. Overreliance could also be exploited by indirect prompt injection, with researchers demonstrating the possibility for a 'hacked' LLM to elicit information from a user [6]. By injecting instructions into an LLM, researchers were able to have the model ask users questions, enabling them to gain information such as the user's real name [6]. If workers over-rely on a generative AI system, they might give over such information in a conversation without thinking of it as being a risk.

The issue of AI literacy, education and equality must be emphasised when integrating generative AI. Once trust and ubiquity of generative AI in business has been built, businesses should consider whether there will be a threat of delegating too much to machines, thereby threatening the *autonomy* of workers. Moreover, what impact could this have to erode the capacity of workers to make choices, especially in significant decision making? Over time, and once AI models become engrained in the operations of a workplace, employee capacity to judge the lines between what the AI can and cannot do might become blurred. For example, is a new staff member, going to understand when to rely on an AI decision and when not to? This will also be a challenge as there is evidence that LLMs change their behaviour over time [82].

We earlier defined justice as it relates to AI as promoting fairness, equality, and shared benefits. As a matter of justice, the displacement of jobs is a recognised threat of AI integration, threatening fairness and equality [46]. Generative AI can be used in internal business process such as human resource management (HRM) for training and development initiatives, resource allocation and employee engagement [83]. But HRM decisions also have an impact on individuals, such as who gets hired or fired, who gets better appraisals, or who is put on preferred projects [84]. These type of decisions all have psychological impacts on employees [85]. If generative AI is used in the process of evaluating staff performance it must be done so in light of distributive justice (everyone is treated the same way by the system) and procedural justice (the processes employed to reach a decision are transparent) [84, 86]. This last point concerns the principle of *explicability*. When implementing a generative AI system, its use and capabilities should be explainable to all users. Management and employees should know why certain systems are used, how they make their decisions and on what information in order to reduce possible overreliance.

7.2 Over-trust

The second factor we identify is what we term as over-trust in generative AI systems. This refers to the degree to which users trust a model with sensitive information, or trust that it is safe and secure. Studies have found that a proportion of employees have pasted sensitive information into ChatGPT [87]. Companies such as Samsung moved to ban employees using ChatGPT as a result of company proprietary material being placed into the program [9]. There is also increasing trust placed in third-party AI providers, without always a consideration of the cyber security risks [12].

Some companies have moved to create their own in-house AI models trained on company data and information to assist staff with queries. BloombergGPT, for example, an LLM that was purpose built from the scratch for finance by Bloomberg [88]. The training and use of such a model brings its own security challenges, as we have seen such models are susceptible to data extraction attacks [3]. Bloomberg, for their part, chose not to release their model citing security concerns of a model trained on so much company data being potentially exposed through nefarious means increasing risk for harm [88]. Training such a model is cost intensive and not something which is an option for many businesses.

Large companies such as Morgan Stanley are using cloud-based systems only accessible to its employees. While some argue a leak of confidential or private information "should not be a problem" [89] this thinking ignores the risk of internal threats and of actors trying to use attacks such as training data extraction. It also ignores the risk of the model being otherwise leaked, as happened with Meta's AI language model LLaMA [90]. There is also the risk of accidental data leaks, such as the recent 38 TB of data accidentally exposed by Microsoft AI researchers [8].

Companies such as Salesforce have touted promises of plugging the AI "trust gap", promoting services to protect company information while using AI tools, a package which will reportedly cost businesses $360,000 per year to implement [91].

With companies implementing domain specific LLMs, in an unregulated market, ethical considerations should still be implemented to protect the security of data. By applying the ethical principles from Sect. 4, we can see how over-trust in a new and untested technology presents ethical issues for companies.

In terms of beneficence, there are many positive benefits for the training of generative AI and large language models on proprietary content or knowledge [89]. This can be useful in assisting customer-facing employees find information about company policy, solving customer problems, or keeping employee knowledge when they leave the organisation [89]. In implementing such a strategy, companies and staff must have best practises in mind, and continually revise its use. Morgan Stanley reportedly used 1,000 financial managers to fine tune its model for safety and use [92]. However, this kind of resource intensive safeguard is not something that is practical for all businesses.

By trusting generative AI systems to store and process data, organisations could also be exposing themselves to added security threats. *Non-maleficence* ("do no harm") in this case does not just mean intentional harm, but also means preventing accidental harm or the harm from the "unpredicted behaviour of machines" [59]. Placing data or sensitive information into generative AI models, could increase the threat of infringing upon personal privacy by increasing a company's exposure to cyber-risk. Generative AI models can be susceptible to attacks such as prompt injection attacks or data extraction attacks, both of which have the potential to leak sensitive data [6]. If we consider that IBM estimates that only 24% of generative AI projects will include a cyber security component within the next 6 months [11], then this rush to adopt AI is leading to users being exposed to unnecessary consequences.

A new question raised by generative AI is what *autonomy* do customers have over their information being stored or used in a model which potentially has flaws in security? If generative AI programs become widespread and ubiquitous in business, should customers have to give their consent for their information to be either (a) be used in the training set of a model; or (b) to be inputted into the finished model?

Regulations about the business use of these models is on the horizon, but there are many questions still to be considered. If users or customers have a right for their data to be erased from a database, such as under the rules of the GDPR, similar protections cannot be offered once a model has been trained a person's data. There is also no option to later withdraw consent once a model has been trained. Mechanisms and best practice around the use of customer information, which could threaten autonomy, must be taken into consideration. The ethical AI guidelines Floridi et al. [59] point out that the autonomy of humans should be promoted, while also limiting the autonomy of machines, and making them intrinsically reversable. The problem with LLMs is that they are lacking in the capacity to be reversable.

Justice in both AI and cyber security encompasses the protection of rights, in particular the right to privacy over data. In using and training models with data taken from users, for example, there must be a consideration for the protection of this data. The susceptibility of models to attacks can include the threat of information or data theft [6].

Justice can also refer to recourse available when something goes wrong with AI systems or in cyber security. As more companies use LLMs the greater the risk becomes of data being leaked. Without clear guidelines or regulation in

place, what recourse do users have if their data is used in a training model and then subsequently exposed? If a company is using a third-party AI provider, is it clear where the responsibility for any failures lies?

A follow on from the ethical considerations of justice, is the principle of explicability. With the rapid implementation of generative AI, are customers being informed whether their data is being used to train new company models? Large companies such as Facebook, Amazon and X (formerly Twitter) all have plans to train LLMs using user data [93]. Amazon plans to train its LLM using voice data from Alexa conversations [93]. Do customers need to opt-in to their data being used to train generative AI models? If their data is exposed in a generative AI hack, who is responsible? What legal avenues could they pursue? Explicability entails who is made accountable for failures in cyber security, in the result of a breach due to generative AI, do companies know who would be at fault or where the responsibility lies?

8 Ethical implementation of generative AI

The above analysis shows the many ethical questions which are raised by thinking about cyber security and generative AI for business. We argue that cyber security needs to be an ethical consideration for businesses implementing generative AI. As such, we offer five key recommendations which companies can adopt to ensure that the security risk of using AI models is limited.

i A secure and ethical AI model design.

When designing an AI model, companies should ensure that their designs take into consideration the principles such as beneficence and non-maleficence. This means considering the potential harms and security risks which could be exposed through the model. Each design should also include non-discriminatory principles to avoid biases and unexpected outcomes from the AI models. Following the principle of explicability, companies should ensure their AI training is easily explainable and transparent in its design.

ii A trusted and fair data collection process.

Companies need to ensure data collected is accurate, fair, representative, and legally sourced. As the principle of autonomy demonstrates, there should be considerations of how much users can have a say about how their data is used in the training of a model. Companies should consider whether they will need to have an opt-in or opt-out systems to protect the privacy of users or customers.

iii A secure data storage.

Companies will need to adhere to the privacy best practices for all the data stored, whether it is training data or input data from users. This should also be done while considering the risk of leaks through hacks such as training data extraction. With regulation of generative AI on the horizon, companies must now prepare by putting in place their own policies over what data is used, while considering the risk that this data could be exposed. This takes into consideration the principle of justice, in the prevention of possible data leaks.

iv Ethical AI model retraining and maintenance.

To maintain model currency and accuracy, AI models require retraining from time to time. Companies need to perform sufficient checks and tests after retraining the AI model and updating the generative AI applications to ensure it maintains its ethical standards and accuracy. In terms of cyber security, this also means constant monitoring for signs of influence, malware or the AI focused attacks as outlined in this paper. New defence training and policies will be needed to monitor for these threats.

v Upskilling, training staff and managing staff.

One of the biggest pain points for business is upskilling and training staff. When implementing a strategy with generative AI, companies should consider what benefit the AI is bringing, while also considering the human impact this will have on staff. If staff are being asked to work with, train or implement models, they might be concerned that they will soon be replaced by these models. Upskilling and training will also be essential to mitigate the potential threats from over-reliance and over-trust in new generative AI models.

9 Conclusion

We have seen that implementation of generative AI comes with considerable cyber security risk for businesses. When rushing to implement generative AI and not fall behind others in industry, companies are also increasing the risk for cyber security breaches. While there is a great momentum toward incorporating generative AI, there also needs to be a

consideration of the ethical responsibility toward the protection of data and prevention against threats.

A major risk with the rush to market of generative AI is its adoption by workers without guidance or understanding of how various generative AI tools are produced, managed or of the risks they pose. This lack of understanding can leave companies open to cyber security threats. We point out two ways in which this can happen: *overreliance* and *over-trust* in generative AI systems. While these two are related, each offers distinct risks and ethical challenges.

The ethical principles of beneficence, non-maleficence, autonomy, justice and explicability are useful lenses through which business can view their obligations when planning to implement data-safe and cyber-secure generative AI solutions.

The rapid adoption of generative AI seems to be moving faster than the industry's understanding of the technology and its inherent ethical and cyber security risks. Companies will need to manage the risk from new vulnerabilities due to generative AI, requiring new forms of governance and regulatory frameworks. Employee training, procedures and managed implementation are an ethical responsibility to protect workers, sensitive company information and the public. Companies now have the opportunity to prevent expensive and unnecessary consequences of generative AI, by addressing the ethical and cyber security threats and investing in data protection measures.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions
Open Access funding enabled and organized by CAUL and its Member Institutions

Declarations

Competing interests On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. IBM: Leadership in the age of AI. IBM: (2023)
2. IBM: The CEO's Guide to Generative AI: Supply chain. IBM: (2023)
3. Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T.B., Song, D.X., Erlingsson, Ú., Oprea, A., Raffel, C.: Extracting Training Data from Large Language Models. In: USENIX Security Symposium. (2020)
4. McKinsey & Company: The Economic Potential of Generative AI: The next Productivity Frontier. McKinsey & Company (2023)
5. Brynjolfsson, E., Li, D., Raymond, L.: Generative AI at Work. National Bureau of Economic Research (2023)
6. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., Fritz, M.: More than you've asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models. arXiv preprint arXiv:2302.12173 (2023)
7. Chui, M., Yee, L., Singla, A., Sukharevsky, A.: The State of AI in 2023: Generative AI's Breakout year. McKinsey & Company (2023)
8. Ben-Sasson, H., Greenberg, R.: 38 TB of data accidentally exposed by Microsoft AI researchers (2023). https://www.wiz.io/blog/38-terabytes-of-private-data-accidentally-exposed-by-microsoft-ai-researchers. Accessed 22 November 2023
9. Park, K.: Samsung bans use of generative AI tools like ChatGPT after April internal data leak (2023). https://techcrunch.com/2023/05/02/samsung-bans-use-of-generative-ai-tools-like-chatgpt-after-april-internal-data-leak/. Accessed 22 November 2023
10. OpenAI: March 20 ChatGPT outage: Here's what happened: (2023). https://openai.com/blog/march-20-chatgpt-outage
11. IBM: The CEO's guide to generative AI: Cybersecurity. IBM: (2023)
12. Renieris, E.M., Kiron, D., Mills, S.: Building Robust RAI Programs as Third-Party AI tools proliferate. MIT Sloan Manage. Rev. (2023)
13. Vallor, S.: An Introduction to Cybersecurity Ethics. Markkula Center for Applied Ethics (2018). https://www.scu.edu/media/ethics-center/technology-ethics/IntroToCybersecurityEthics.pdf
14. Formosa, P., Wilson, M., Richards, D.: A principlist framework for cybersecurity ethics. Computers Secur. **109**, 102382 (2021). https://doi.org/10.1016/j.cose.2021.102382
15. Blanken-Webb, J., Palmer, I., Campbell, R.H., Burbules, N.C., Bashir, M.: Cybersecurity Ethics. Foundations of Information Ethics, pp. 91–101. American Library Association (2019)
16. Morgan, G., Gordijn, B.: A care-based stakeholder approach to ethics of cybersecurity in business. In: Christen, M., Gordijn, B., Loi, M. (eds.) The ethics of cybersecurity https://doi.org/ (2020). https://doi.org/10.1007/978-3-030-29053-5_6, pp. 119–138
17. Agrafiotis, I., Nurse, J.R.C., Goldsmith, M., Creese, S., Upton, D.: A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate. J. Cybersecur. **4** (2018). https://doi.org/10.1093/cybsec/tyy006
18. IBM: Cost of a Data Breach Report 2023. IBM: (2023)
19. Schatz, D., Bashroush, R., Wall, J.: Towards a more representative definition of Cyber Security. J. Digit. Forensics Se. **12**, 53–74 (2017)
20. National Institute of Standards and Technology:, https://csrc.nist.gov/glossary/term/integrity
21. Manjikian, M.: Cybersecurity Ethics: An Introduction. Routledge, London (2023)
22. Christen, M., Gordijn, B., Loi, M.: The Ethics of Cybersecurity. The International Library of Ethics. Law Technol. (2020). https://doi.org/10.1007/978-3-030-29053-5
23. Finlay, C.J.: Just War, Cyber War, and the Concept of Violence. Philos. Technol. **31**, 357–377 (2018). https://doi.org/10.1007/s13347-017-0299-6

24. Taddeo, M.: Information Warfare: A Philosophical Perspective. The Ethics of Information Technologies10.4324/9781003075011-35, pp. 461–476. Routledge (2020)
25. Taddeo, M.: An analysis for a just cyber warfare. 4th Int. Conf. Cyber Confl. (CYCON 2012). **pp 1–10**, Tallinn–Estonia (2012). (2012)
26. Macnish, K., van der Ham, J.: Ethics in cybersecurity research and practice. Technol. Soc. **63** (2020). https://doi.org/10.1016/j.techsoc.2020.101382
27. Van De Poel, I.: Core Values and Value Conflicts in Cybersecurity: Beyond Privacy Versus Security, pp. 45–71. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-29053-5_3
28. Jaquet-Chiffelle, D.-O., Loi, M.: Ethical and Unethical Hacking, pp. 179–204. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-29053-5_9
29. Brey, P.: Ethical Aspects of Information Security and Privacy, pp. 21–36. Springer, Berlin Heidelberg (2007). https://doi.org/10.1007/978-3-540-69861-6_3
30. Riley, S.: DarkLight Offers First of its Kind Artificial Intelligence to Enhance Cybersecurity Defenses. (2017). https://www.businesswire.com/news/home/20170726005117/en/DarkLight-OffersKind-Artificial-Intelligence-Enhance-Cybersecurity. Business Wire Accessed 05 February 2024
31. Li, J.H.: Cyber security meets artificial intelligence: A survey. Front. Inf. Tech. El. **19**, 1462–1474 (2018). https://doi.org/10.1631/Fitee.1800573
32. Lecun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature. **521**, 436–444 (2015). https://doi.org/10.1038/nature14539
33. Kumar, S., Gupta, U., Singh, A.K., Singh, A.K.: Artificial Intelligence: Revolutionizing Cyber Security in the Digital era. J. Computers Mech. Manage. **2**, 31–42 (2023). https://doi.org/10.57159/gadl.jcmm.2.3.23064
34. Carlini, N., Jagielski, M., Choquette-Choo, C.A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., Tramèr, F.: Poisoning Web-Scale Training Datasets is Practical. (2023). ArXiv abs/2302.10149
35. Foster, D.: Generative deep Learning. O'Reilly Media, Inc. (2022)
36. Sallam, M.: In: Healthcare (ed.) ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns, p. 887. MDPI (2023)
37. Cascella, M., Montomoli, J., Bellini, V., Bignami, E.: Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios. J. Med. Syst. **47**, 33 (2023)
38. Lo, C.K.: What is the impact of ChatGPT on education? A rapid review of the literature. Educ. Sci. **13**, 410 (2023)
39. Stokel-Walker, C.: ChatGPT listed as author on research papers: Many scientists disapprove. Nature. **613**, 620–621 (2023)
40. Hutson, J., Harper-Nichols, M.: Generative AI and Algorithmic Art: Disrupting the Framing of Meaning and Rethinking the Subject-Object Dilemma. Global Journal of Computer Science and Technology: D 23, (2023)
41. Pavlik, J.V.: Collaborating with ChatGPT: Considering the implications of generative artificial intelligence for journalism and media education. Journalism Mass. Communication Educ. **78**, 84–93 (2023)
42. Dell'Acqua, F., McFowland, E., Mollick, E.R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., Lakhani, K.R.: Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Harvard Business School Technology & Operations Mgt. Unit Working Paper (2023)
43. Chen, B., Wu, Z., Zhao, R.: From fiction to fact: The growing role of generative AI in business and finance. J. Chin. Economic Bus. Stud. **21**, 471–496 (2023). https://doi.org/10.1080/14765284.2023.2245279
44. Wamba, S.F., Queiroz, M.M., Jabbour, C.J.C., Shi, C.V.: Are both generative AI and ChatGPT game changers for 21st-Century operations and supply chain excellence? Int. J. Prod. Econ. **265**, 109015 (2023)
45. Stahl, B.C., Eke, D.: The ethics of ChatGPT–Exploring the ethical issues of an emerging technology. Int. J. Inf. Manag. **74**, 102700 (2024)
46. Krzysztof Wach, C.D.D., Joanna Ejdys, R., Kazlauskaitė, P., Korzynski, G., Mazurek: Joanna Paliszkiewicz, Ewa Ziemba: The dark side of Generative Artificial Intelligence: A critical analysis of controversies and risks of ChatGPT
47. Zarifhonarvar, A.: Economics of chatgpt: A labor market view on the occupational impact of artificial intelligence. J. Electron. Bus. Digit. Econ. (2023)
48. Gross, N.: What chatGPT tells us about gender: A cautionary tale about performativity and gender biases in AI. Social Sci. **12**, 435 (2023)
49. Ray, P.P.: ChatGPT: A Comprehensive Review on Background, Applications, key Challenges, bias, Ethics, Limitations and Future Scope. Internet of Things and Cyber-Physical Systems (2023)
50. Rahman, M.M., Watanobe, Y.: ChatGPT for Education and Research: Opportunities, threats, and strategies. Appl. Sci. **13**, 5783 (2023), https://doi.org/10.3390/app13095783
51. De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G.P., Ferragina, P., Tozzi, A.E., Rizzo, C.: ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health. Front. Public. Health. **11**, 1166120 (2023)
52. Ferrara, E.: Social bot detection in the age of ChatGPT: Challenges and opportunities. First Monday (2023)
53. OpenAI: GPT-4 System Card. OpenAI: (2023)
54. Fabian, D., Crisp, J.: Why Red Teams Play a Central Role in Helping Organizations Secure AI Systems. Google (2023)
55. Sebastian, G.: Do ChatGPT and other AI Chatbots pose a cybersecurity risk? Int. J. Secur. Priv. Pervasive Comput. **15**, 1–11 (2023). https://doi.org/10.4018/ijsppc.320225
56. Gupta, M., Akiri, C., Aryal, K., Parker, E., Praharaj, L.: From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy. IEEE Access. (2023)
57. Schlagwein, D., Willcocks, L.: ChatGPT et al.': The ethics of using (generative) artificial intelligence in research and science. J. Inform. Technol. **38**, 232–238 (2023). https://doi.org/10.1177/02683962231200411
58. Illia, L., Colleoni, E., Zyglidopoulos, S.: Ethical implications of text generation in the age of artificial intelligence. Bus. Ethics Environ. Responsib. **32**, 201–210 (2023). https://doi.org/10.1111/beer.12479
59. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., Vayena, E.: AI4People—An ethical Framework for a good AI society: Opportunities, risks, principles, and recommendations. Mind. Mach. **28**, 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5
60. Bruschi, D., Diomede, N.: A framework for assessing AI ethics with applications to cybersecurity. AI Ethics. **3**, 65–72 (2023). https://doi.org/10.1007/s43681-022-00162-8
61. Van De Poel, I.: An ethical Framework for evaluating Experimental Technology. Sci Eng. Ethics. **22**, 667–686 (2016). https://doi.org/10.1007/s11948-015-9724-3
62. Hosseini, Z., Nyholm, S., Le Blanc, P.M., Preenen, P.T.Y., Demerouti, E.: Assessing the artificially intelligent workplace: An ethical framework for evaluating experimental technologies in workplace settings. AI Ethics. (2023). https://doi.org/10.1007/s43681-023-00265-w

63. Himma, K.E.: The Ethics of tracing Hacker attacks through the machines of innocent persons. Int. Rev. Inform. Ethics. **2** (2004). https://doi.org/10.29173/irie256
64. Franceschelli, G., Musolesi, M.: Copyright in generative deep learning. Data Policy **4**, e17 (2022)
65. Kirk, H.R., Jun, Y., Volpin, F., Iqbal, H., Benussi, E., Dreyer, F., Shtedritski, A., Asano, Y.: Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. Adv. Neural. Inf. Process. Syst. **34**, 2611–2624 (2021)
66. Spinello, R.A.: Corporate Data breaches: A Moral and Legal Analysis. J. Inform. Ethics. **30**, 12–32 (2021). https://doi.org/https://doi.org/10.2307/JIE.30.1.12
67. Erzberger, A.: WormGPT and FraudGPT – The Rise of Malicious LLMs (2023). https://www.trustwave.com/en-us/resources/blogs/spiderlabs-blog/wormgpt-and-fraudgpt-the-rise-of-malicious-llms/. Accessed 27 November 2023
68. Group-IB: Group-IB Discovers 100K + Compromised ChatGPT Accounts on Dark Web Marketplaces: (2023). https://www.group-ib.com/media-center/press-releases/stealers-chatgpt-credentials/. Accessed 27 November 2023
69. OpenAI: Introducing GPTs: (2023). https://openai.com/blog/introducing-gpts
70. Gelper, S., van der Lans, R., van Bruggen, G.: Competition for attention in online social networks: Implications for seeding strategies. Manage. Sci. **67**, 1026–1047 (2021)
71. Caramancion, K.M.: An exploration of disinformation as a cybersecurity threat. In: 2020 3rd International Conference on Information and Computer Technologies (ICICT), pp. 440–444. IEEE, (2020)
72. Petratos, P.N., Faccia, A.: Fake news, misinformation, disinformation and supply chain risks and disruptions: Risk management and resilience using blockchain. Ann. Oper. Res. **327**, 735–762 (2023). https://doi.org/10.1007/s10479-023-05242-4
73. Petratos, P.N.: Misinformation, disinformation, and fake news: Cyber risks to business. Bus. Horiz. **64**, 763–774 (2021)
74. Goldstein, J.A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., Sedova, K.: Generative language models and automated influence operations: Emerging threats and potential mitigations. arXiv Preprint arXiv:230104246 (2023)
75. Edwards, B.: AI-powered Bing Chat spills its secrets via prompt injection attack (2023). https://arstechnica.com/information-technology/2023/02/ai-powered-bing-chat-spills-its-secrets-via-prompt-injection-attack/. Accessed 27 November 2023
76. Boxleitner, A.: Pushing Boundaries or Crossing Lines? The Complex Ethics of ChatGPT Jailbreaking. The Complex Ethics of ChatGPT JailbreakingOctober 17, (2023) (2023)
77. Buçinca, Z., B Malaya, M., Z Gajos, K.: To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proc. ACM Hum. -Comput Interact. **5**, Article188 (2021). https://doi.org/10.1145/3449287
78. Vasconcelos, H., Jörke, M., Grunde-Mclaughlin, M., Gerstenberg, T., Bernstein, M.S., Krishna, R.: Explanations Can Reduce Overreliance on AI Systems During Decision-Making. Proceedings of the ACM on Human-Computer Interaction 7, 1–38 (2023). https://doi.org/10.1145/3579605
79. Skitka, L.J., Mosier, K.L., Burdick, M.: Does automation bias decision-making? Int. J. Hum-Comput St. **51**, 991–1006 (1999). https://doi.org/10.1006/ijhc.1999.0252
80. Cummings, M.: Automation bias in intelligent time critical decision support systems. In: AIAA 1st intelligent systems technical conference, pp. 6313. (2004)
81. Pearce, H., Ahmad, B., Tan, B., Dolan-Gavitt, B., Karri, R.: Asleep at the keyboard? assessing the security of github copilot's code contributions. In: 2022 IEEE Symposium on Security and Privacy (SP), pp. 754–768. IEEE, (2022)
82. Chen, L., Zaharia, M., Zou, J.: How is ChatGPT's behavior changing over time? arXiv preprint arXiv:2307.09009 (2023)
83. Ooi, K.-B., Tan, G.W.-H., Al-Emran, M., Al-Sharafi, M.A., Capatina, A., Chakraborty, A., Dwivedi, Y.K., Huang, T.-L., Kar, A.K., Lee, V.-H., Loh, X.-M., Micu, A., Mikalef, P., Mogaji, E., Pandey, N., Raman, R., Rana, N.P., Sarker, P., Sharma, A., Teng, C.-I., Wamba, S.F., Wong, L.-W.: The potential of Generative Artificial Intelligence Across disciplines: Perspectives and future directions. J. Comput. Inform. Syst. 1–32 (2023). https://doi.org/10.1080/08874417.2023.2261010
84. Tambe, P., Cappelli, P., Yakubovich, V.: Artificial intelligence in human resources management: Challenges and a path forward. Calif. Manag. Rev. **61**, 15–42 (2019)
85. Varma, A., Dawkins, C., Chaudhuri, K.: Artificial intelligence and people management: A critical assessment through the ethical lens. Hum. Resource Manage. Rev. **33**, 100923 (2023)
86. Robert, L.P., Pierce, C., Marquis, L., Kim, S., Alahmad, R.: Designing fair AI for managing employees in organizations: A review, critique, and design agenda. Human–Computer Interact. **35**, 545–575 (2020). https://doi.org/10.1080/07370024.2020.1735391
87. Cameron, C.: 11% of data employees paste into ChatGPT is confidential (2023). https://www.cyberhaven.com/blog/4-2-of-workers-have-pasted-company-data-into-chatgpt. Accessed 23 November 2023
88. Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., Mann, G.: Bloomberggpt: A large language model for finance. arXiv Preprint arXiv:230317564 (2023)
89. Davenport, T., Alavi, M.: How to Train Generative AI Using Your Company's Data (2023). https://hbr.org/2023/07/how-to-train-generative-ai-using-your-companys-data. Accessed 27 November 2023
90. Vincent, J.: Meta's powerful AI language model has leaked online — what happens now? (2023). https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse. Accessed 27 November 2023
91. Lin, B., Loten, A.: Salesforce Aims to Plug 'AI Trust Gap' With New Tech Tools (2023). https://www.wsj.com/articles/salesforce-aims-to-plug-ai-trust-gap-with-new-tech-tools-19e11750. Accessed 27 November 2023
92. Bautzer, T., Nguyen, L.: Morgan Stanley to launch AI chatbot to woo wealthy (2023). https://www.reuters.com/technology/morgan-stanley-launch-ai-chatbot-woo-wealthy-2023-09-07/. Accessed 27 November 2023
93. Leffer, L., Your Personal Information Is Probably Being Used to Train Generative AI Models: (2023). https://www.scientificamerican.com/article/your-personal-information-is-probably-being-used-to-train-generative-ai-models/. Accessed 27 November 2023