Socio-Economic Threats of Deepfakes and The Role of Cyber-Wellness Education in Defense

MILAD TALEBY AHVANOOEY*, Warsaw University of Technology, Poland and Nanyang Technological University, Singapore
WOJCIECH MAZURCZYK, Warsaw University of Technology, Poland
DONGWON LEE, The Pennsylvania State University, USA

In recent years, society has witnessed accelerated advancement in generative artificial intelligence (GenAI) technologies, which may be viewed as a double-edged sword. On one hand, GenAI tools can be used to create synthetic content legitimately. For example, advertising agencies may, with permission, generate celebrities' images or videos using GenAI tools without putting them in front of cameras and thus reducing the overall cost of media construction. On the other end, scammers may utilize GenAI tools to craft or edit artificial contents (e.g., texts, images, videos, or audios), so-called 'deepfake,' to mislead or deceive netizens, i.e., robocalls or voice cloning phishing, potentially causing detrimental consequences for society. This paper briefly debates emerging socio-economic threats of deepfakes in today's society and how cyber-wellness (or digital media literacy) education can help netizens mitigate their risks.

CCS Concepts: • **Security and privacy → Social aspects of security and privacy**; **Social network security and privacy**.

Additional Key Words and Phrases: GenAI risks, Deepfake threats, Cyber-wellness education, digital media literacy

ACM Reference Format:
Milad Taleby Ahvanooey, Wojciech Mazurczyk, and Dongwon Lee. 2025. Socio-Economic Threats of Deepfakes and The Role of Cyber-Wellness Education in Defense. In *Proceedings of Communications of the ACM*. ACM, New York, NY, USA, Article 1, 14 pages. https://doi.org/10.1145/3715317

1 Introduction

Due to the limits of science and its learnable dependencies (e.g., time and cost), we must rely on the expertise of others to form our knowledge and skills [28]. For example, social media platforms have drastically revolutionized how netizens (i.e., users who are actively engaged in online communities) learn knowledge and skills by exchanging costless information with the public (e.g., followers or influencers). Globally, business owners and malicious actors also use such platforms and tools based on generative artificial intelligence (GenAI) to craft or edit synthetic media to expand their revenue by attracting more customers and improving the experiences of their target consumers [30]. GenAI-crafted content in the form of various media called deepfake[1], is digitally manipulated

**Corresponding Author

[1] https://www.oed.com/dictionary/deepfake_n

Authors' Contact Information: Milad Taleby Ahvanooey, Warsaw University of Technology, Warsaw, Mazowieckie, Poland and Nanyang Technological University, Singapore, Singapore, M.taleby@ieee.org; Wojciech Mazurczyk, Warsaw University of Technology, Warsaw, Mazowieckie, Poland, wojciech.mazurczyk@pw.edu.pl; Dongwon Lee, The Pennsylvania State University, University Park, PA, USA, Dongwon@psu.edu.

Communications of the ACM, 2025

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/3715317

to resonate one thing's or person's likeness conclusively with that of another, and it is often utilized maliciously to manifest an event that has not happened in reality. Such synthetic media may contain partial facts to mislead viewers (called **misinformation)**. In a recent study conducted by iProov [16], consumers ($n = 16,000$) participated in a survey across eight countries in 2022 where only 71% of global respondents knew "what a deepfake video is", and 57% recognized the difference between a real video and an artificially crafted video. According to a recent poll report published by the New York Post on March 13, 2024 [26], the survey of 2,000 registered voters in the USA showed not only that participants are increasingly pessimistic concerning a political campaign advertised by deepfakes, but also they could not distinguish between artificial contents and human-created ones, i.e., approximately 43% or half of those believe GenAI generated contents negatively affect the result of 2024 elections.

In June 2023, to highlight the danger of mass-crafted spurious information, a group of two engineers developed an experimental "propaganda machine" called *CounterCloud*[2] using the open-source GenAI models. In this exemplary test case of the project, the crafted deepfake texts [1], [35] (e.g., 50 tweets and 20 news articles daily) were seemingly convincing 90% of the time. The primary goal behind this project was to enlighten people about how unbelievably effortless it is to weaponize GenAI to propagate deliberately manipulated contents called **disinformation** on a global scale. In addition, the creators claimed that with approximately $4K monthly cost, anyone can produce more than 200 articles per day. This may counter more than 40 news outlets without requiring any human interaction, and this can be enough to, for example, influence an election campaign. In addition to disinformation and misinformation, there are **malinformation** forms of deepfake threats [32], which are based on manipulated facts to harm targeted victims (e.g., cyber-scamming [20] and cyberbullying [27]). Therefore, netizens must learn of any suspicious activities where malicious actors may use malinformation to approach them through hidden traps in decision-making by crafting and applying deepfakes to benefit from their lack of knowledge/awareness.

2 Socio-economic Threats of Deepfakes

As evidenced within recent cases reported by mass media and law enforcement agencies [27], scammers create and deploy deepfake content (e.g., voice and/or video) that poses as a colleague, family member, or a celebrity known for the target netizens to deceive them. When netizens lack the necessary awareness, knowledge, and skills to identify deepfakes, they may eagerly trust such synthetic contents and likely be fooled, leading to unwise actions or/and financial losses. For example, in a recent cyber-scamming incident using GenAI tools in February 2024, $25.6 million was stolen from a multinational firm in Hong Kong by tricking an employee into believing that the CEO requested the fund transfer via a video call [7]. Therefore, if the victim clicks on the link connected to the deepfake ads and transfers some money or cryptocurrency, (s)he may fall into a scammer's phishing trap [20] and probably face financial loss. Recently, in a consumer alert recommended by the US Federal Trade Commission in March 2023 [29], an education specialist stated that voice-cloning scams in the guise of family emergency calls are rising dramatically. To gain the victim's trust, a scammer simply needs to find a short recorded voice of his/her family member, which could often be available on their social media profile, and then apply a GenAI tool (e.g., LivePerson voice AI chatbot) to create a deepfake voice conversation by following nefarious purposes [13], which is virtually indistinguishable from the actual one.

To avoid a voice cloning trap, for instance, the victim should not trust him/her, and instead (s)he must apply cyber-wellness knowledge to call the claimed person and ask verifiable questions to confirm his/her authenticity [27]. In another form of cyber-scams, crafted deepfake videos of

[2] https://www.youtube.com/watch?v=cwGdkrc9i2Y

celebrities were widely spread on social media platforms by promoting services (e.g., the Quantum AI Elon Musk trading bot [20]) linked to phishing websites. Moreover, in a public alert in June 2023 [27], the US Federal Bureau of Investigation (FBI) announced that they received reports on cyber criminals deploying GenAI tools to craft artificial porn videos or photos of underage victims (e.g., minor children). The offenders then sent the generated synthetic contents directly to victims for cyberbullying or sextortion purposes. For instance, to combat such unprecedented threats, the US National Center for Missing and Exploited Children has built an online platform (called "Take It Down"[3]), which provides free service to help victims facing online nudity incidents to stop and prevent the spread of videos or images, particularly sexually explicit contents of underage children. In general, netizens who apply GenAI tools for various purposes can act as unintentional propagators of deepfakes due to a lack of proper awareness, knowledge, and skills in writing well-structured promotes and validating the reliability and accuracy of the outcomes.

These cyberthreats are an ongoing worldwide phenomenon with significant implications in many aspects of economics and society, such as cryptomarket trading, elections, health, and education. In a recent regulatory action in February 2024 [9], the USA Federal Communications Commission (FCC) declared that unwanted GenAI-crafted robocalls and robotexts should be banned by laws, as well as closed a comment period on public citizen's petition for regulating a new rule concerning the application of deepfakes in election advertisements. This legislation action was initiated after the propagation of a robocall on the 21st of January 2024, in which President Joe Biden discouraged citizens from voting. In addition, they expect that once this petition is approved at the national level, such regulations should be recommended internationally under the aegises of organizations such as the European Union and United Nations, as well as tech-related consortiums like the Christchurch Call[4].

3 Educational Aspects

While the use of GenAI tools could be justified in some cases (e.g., licensed GenAI-based commercials), the dangers of these technologies are beyond anyone's imagination. For example, netizens can easily access GenAI software for free or pay a membership subscription for various applications (see Table 1). On the other hand, the GenAI tools provide opportunities to innovate and reform many industries (e.g., education and advertisement). Still, they can negatively impact the lifelong learning of netizens in society and have severe consequences on developing their critical and creative thinking skills. A recent experimental study showed that the OpenAI GPT-4 could gain considerable scores on standardized tests, such as 99% on GRE Verbal and 89% on SAT Math [24] due to applying more collaborative, creative, and efficient transformers compared to previous versions. Moreover, a survey study demonstrated that over 51% of students believed applying GenAI software (e.g., ChatGPT) to pass exams or prepare multimedia assignments is technically cheating [14]. For instance, the GPT-4o is the latest flagship model of OpenAI ChatGPT, which offers the same level of intelligence as the GPT-4, but it is much faster and enhances its capabilities across text, vision, and voice. Evidently, public usage of such "unaccountable and unexplainable" GenAI tools raises societal and governmental concerns about why they are easily accessible to everyone. For instance, several countries recently banned ChatGPT usage in public universities by blocking access to it through the Internet networks inside schools. Furthermore, they decided that using GenAI tools to prepare assignments would be considered academic misconduct. This includes, for example, at least five Australian states and eight elite universities in the Russell Group in the UK, consisting of Oxford and Cambridge [5]. Since GenAI tools are rapidly evolving, crafted outcomes are becoming

[3] https://takeitdown.ncmec.org/

[4] https://www.christchurchcall.org/

practically impossible to distinguish from actual human-generated media. Consequently, regulatory agencies and industrial sectors must jointly develop educational programs to guide netizens in making safe decisions when dealing with deepfake contents. In an exemplary policy action, the Data Protection Agency of Italy temporarily restricted ChatGPT's services in this country. Subsequently, to incorporate this criticism, OpenAI agreed to perform a series of updates to its online privacy policies and notices, such as optionality, security, and transparency [2]. However, these regulatory policies do not provide sufficient guidelines for netizens to learn how to decide and act safely when facing doubtful media.

TABLE 1. An overview of GenAI tools and their weaponized threats.

Summary of GenAI Model Types	GenAI Tools	Target Use Cases (+) and Weaponized Threats (-)
LLMs can craft or edit textual contents based on a prompt.	ChatGPT, Brad, and ChatSonic;	+ Creative business, and academic writing; + Source code generation for programming; + Textual content translation; + Realtime chatbots to create robotexts; - Misinformation, disinformation, and malinformation; - Copyright and ownership violation of copyrighted texts; - Reeducation of creative thinking in netizens;
VLMs can alter or create images based on a prompt.	Midjourney, DALL-E, Jasper, and Stable diffusion;	+ Marketing, blogging, and other purposes; + Automatic image editing and retouching; - Cyber-scamming, Cyberbullying & sextortion; - Copyright and ownership violation of copyrighted images;
MLLMs can compose or edit songs (e.g., lyrics and melody) based on a prompt.	Jukebox, Bloomy AI, Splash Pro, and Magenta Studio;	+ Music composing and editing; + Songwriting and rhyming song lyrics; - Copyright and ownership violation of copyrighted musics; - Misinformation, disinformation, and malinformation;
LMMs can craft or edit videos or any other multimedia contents based on a prompt.	Runway Gen-2, VEED.IO, Colossyan creator, Synthesia AI, Runway, Luma AI (Genie), Masterpiece Studio, Get3D, and Spline AI;	+ Movie content creation; + Animation and 3D models generation; + Game design and development; - Cyber-scamming, cyberbullying & sextortion; - Copyright and ownership violation of copyrighted videos; - Misinformation, disinformation, and malinformation;
SLLMs can create or edit speeches based on a prompt.	ChatGPT-4o, Google Translate, Baidu Translate, Microsoft Translator, LivePerson, Murf AI, WaveNet, and Lovo AI;	+ Audiobook creation; + Dubbing and speech generation for accessibility; + Real-time voice translation; + Voice chatbots; - Voice cloning robocalls; - Copyright and ownership violation of copyrighted voices; - Misinformation, disinformation, and malinformation;

* Note that a prompt is a conceptual means of instruction(s), which netizens must input to guide the GenAI tool for constructing or editing contents, whether it is normally a word-based text, image, audio, or a combination of these media.

The science of **Cyber-Wellness Education** (CWE) involves teaching standard guidelines to netizens to understand preventive mechanisms and sufficient awareness of how to protect and stay safe while interacting with cyberspace [17]. Over the last decades, the European Union[5] and many other countries (e.g., USA[6], and UK[7]) have developed digital literacy and/or CWE programs (see Table 2) and are actively using them in educational institutes to reduce unprecedented risks of cyberthreats. Numerous successful online scams have recently caused netizens to lose millions of dollars [20]. This is mainly because malicious actors apply various forms of deepfakes to devise unprecedented cyberthreats, and the current CWE programs are not sufficiently updated, making them inadequate for netizens even if they get trained by such curricula. Moreover, these programs focus on training students and pay less attention to other regular consumers of digital contents on the Internet.

[5] https://www.digital-wellbeing.eu
[6] https://lincs.ed.gov/
[7] https://www.gov.uk/

Table 2. Existing CWE (or digital media literacy) programs and governmental action plans.

Action Plans	Details	List of Contents
Digital-Wellbeing Education[5] (Co-founded by the EU Programme and Erasmus)	This is a curriculum that consists of CWE course materials, which are aimed for trainers and educators to deliver knowledge and skills as part of their digital media literacy programs or to update or integrate an existing program. Such materials provide instructors with up-to-date resources and practical knowledge, and skills to guide and train them to ensure their students are educated in digital wellbeing.	- Introduction to Digital Wellbeing- Self-Image - Online and Offline Identities - Digital Footprint, Netiquette, and Reputation - Cyber Bullying and Conflict Resolution - Privacy, Security, and Safety - Personal Goals and Managing Distractions - Ultimate Guide to Creating a Professional LinkedIn - Critical Thinking, Fake News, and Extreme Views - Digital Citizenship and Social Responsibility
Teaching Skills that Matter: Digital Literacy[6] (Founded by the USA's Department of Education)	The USA's LINCS system provides a set of resources on digital literacy to offer best practices, lesson plans on social media platforms and workplace safety, and two types of learning templates (problem- and project-based). Such resources enable netizens to find, assess, construct, and communicate information; and form digital citizenship and practice responsible usage of technologies.	- Digital Literacy: Issue Brief - Best Practices in Digital Literacy: A Case Study - Social Media Lesson Plan - Workplace Safety Lesson Plan - Sharing Information about Important Safety Signs-Integrated and Contextualized Learning Lesson - Cultural Stereotypes Online Problem-based Learning Lesson - Folk Stories Project-based Learning Lesson - Annotated Instructional Resources and References - Teaching the Skills That Matter: Digital Literacy in Action
Online Media Literacy Strategy[7] (Founded by the UK Department for science, Innovation, and Technology, and Department for Digital, Culture, Media & Sport)	This programme aims to train and empower netizens across the UK to control their safety through online platforms. Over 170 organizations are presently involved in delivering CWE in the UK. This strategy outlines the government's plan to coordinate media literacy landscape in several years and provides a CWE framework for the best practical principles to apprise the contents and delivery of up-to-date educational materials.	- Data and privacy - Online environment - Information consumption - Online consequences - Online engagement

As outlined in Table 1, GenAI tools and social media platforms provide opportunities for netizens to craft and spread propaganda and manipulate information in an easy manner. These contents are intentionally crafted deepfakes to deceive their target audiences. In addition to such intentional deepfakes, traditional media outlets, online libraries, and social media platforms can be both perpetrators or victims of sharing unintentional fabricated contents. When it comes to crafting content using GenAI tools, there are situations where netizens write a prompt, over-trust the result, and use it unintentionally for sensitive purposes [32], which is entirely wrong and must be validated due to possibilities of underpinning models' biases and hallucinations [30]. For instance, a New York attorney deployed the ChatGPT to conduct legal research to represent a client's injury claim in May 2023. While overseeing the suit, the federal judge noticed that six citations quoted in the attorney's brief were falsified[8]. In this hallucination scenario, the ChatGPT made cited cases up and even emphasized realistically that they were accessible in major legal databases.

4 GenAI Distortion Risks in Society

GenAI tools' tendency to craft seemingly realistic media by combining facts with fiction may often lead to spurious information (or invalid concepts) that distort netizens' perceptions of contents' reality and their associated dangers – a phenomenon called *GenAI distortion risks* in society [38]. One way to reduce such risks is to learn how to apply prompting knowledge and skills correctly as well as understand decision-making biases that help netizens design and refine effective prompts for crafting more accurate outcomes from GenAI tools and consuming deepfakes [36]. Technically, this is the process of crafting or refining a query: one or more connected sentences given to GenAI tools that may result in producing more accurate and relevant contents. In other words, the output validity and quality of the GenAI tools are influenced by two independent factors, the lack of which can negatively change their accuracy [18]. These factors include *i)* Coherence: the quality of a prompt being logically close to what a user expects to gain as content, and *ii)* Relevance: the

[8] https://www.cnn.com/2023/05/27/business/chat-gpt-avianca-mata-lawyers

availability of associated data to be processed and collected by web scraping approaches through GenAI tools. Indeed, the enhancement of netizens' knowledge and skills on how to craft and refine prompts can lead to a more accurate conceptualization and, eventually, production of the contents by the Large Language Models (LLMs), Large Vision Language Models (VLMs), Music-specialized Large Language Models (MLLMs), Speech Large Language Models (SLLMs), or Large Multimodal Models (LMMs). Since GenAI tools (see Table 1) deploy web scraping techniques to collect relevant data from different accessible resources on the Internet, they are somewhat limited. Therefore, if there is insufficient or incorrect associated data related to the prompts to be fed to the GenAI models, they may produce biased content containing irrelevant concepts. Technically, these biased outcomes are caused by three **data-driven biases** [37]: i) problematic training data is an unintentional perpetuation of bias that may link inaccurate or incorrect data to specific subjects; ii) the accessibility to real-time data is a significant limitation, which can cause the recency bias; and iii) underpinning models can be biased by design and then embed biases into the associated data that inevitably produce unfair outcomes [40]. In addition to data-driven biases, there are three **decision-making biases** that can be caused by netizens: i) the tendency to over-trust AI tools, which leads to a false confirmation; ii) optimizing the prompt might steer the GenAI models to adapt their replies toward netizen's objectives that can cause feedback loop bias [30]; and iii) the act of using a GenAI-crafted content for commercial purposes (e.g., journalism or scientific publications) might violate the ownership right of similar contents that can be interpreted as anti-copyright bias.

Practically, the above-mentioned data-driven biases may cause GenAI hallucinations that result in unintentional deepfakes if the netizens trust such contents while dealing with decision-making biases. To reduce the possibility of creating invalid contents using GenAI tools, we discuss the necessary knowledge and skills that guide netizens to understand the validity and reliability of GenAI-crafted information by following a prompting protocol. Indeed, the process of crafting content requires well-structured human-GenAI interventions to operate accurately and achieve high-standard results. Hence, the following five elements must be considered when a user writes an effective prompt.

1) Relevance: An effective prompt must be relevant to the expected task to be done by providing adequate information to direct the GenAI tool to make precise predictions [22].

2) Diversity: An effective prompt should contain a range of specific details to ensure the generalization of the integrated GenAI models to new data [8].

3) Consistency: An effective prompt should include a consistent format and corrected format to feed the GenAI models to learn the requested tasks effectively [8].

4) Simplicity: An effective prompt can be concise and uncovered to reduce the possibilities of confusion for GenAI models when processing the requested tasks [30].

5) Clarity: An effective prompt should be clearly defined unambiguously so that the task concept meets the best possible quality of transparency that the GenAI models can perform.

Indeed, effective prompts can result in more informative and accurate GenAI outcomes, while defectively designed ones may construct irrelevant and confusing responses [31]. In addition to considering the above key features, netizens must learn that using GenAI services responsibly is crucial if they wish to apply the full capabilities of such tools without forfeiting their integrity and ingenuity. Since GenAI tools technically generate contents by interpreting a given prompt and reproducing new information based on the trained data accordingly, there is a possibility they may construct incorrect or misleading results, known as "hallucinations" [25]. This could be a common problem through GenAI tools, as they are developed to craft new forms of content and can sometimes create plausible-sounding content, which in the end might be incorrect information [15]. Surprisingly, most corporate systems (or service providers) are aware of such constraints and are gradually advancing and optimizing their GenAI models to enhance the accuracy of their outputs as much as possible. Since the capability of GenAI tools depends on interpreting netizens' prompts, they should be assumed collaborative entities in the knowledge co-construction process [34]. To enrich the human-GenAI knowledge co-construction, we extended the prompting protocol introduced by Robertson et al. [30]. It helps netizens to write more detailed and effective prompts, which eventually brings out the best capability of GenAI tools as well as a question-based framework to validate their accuracy.

In the constructivist theory, learners (hereafter referred to as netizens) usually construct knowledge instead of passively imitating information. Commonly, they experience the environment around them and reflect upon their observations, and eventually form their perceptions by interpreting and incorporating new concepts into their pre-existing information. Similarly, when given a prompt, a GenAI tool crafts relevant content by predicting the new concept through the interpretation process [30]. In the context of human-GenAI prompting, performing an iterative refinement process is necessary to improve the dialogue between the GenAI tools and netizens, formulating a cohesive reply to an inquiry as a problematic task. Hence, considering the above constructivist perspectives [30], the prompting process can be formed as an iterative and circular-based protocol with three correlated dimensions and nine steps that can coherently facilitate knowledge co-construction and its validation by empowering human-GenAI dialogue (see Figure 1).

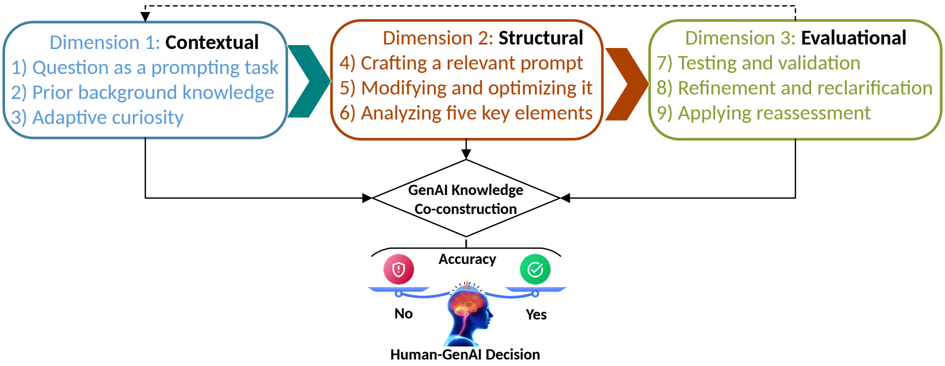
Dimension 1: Contextual
1) Question as a prompting task
2) Prior background knowledge
3) Adaptive curiosity
Dimension 2: Structural
4) Crafting a relevant prompt
5) Modifying and optimizing it
6) Analyzing five key elements
Dimension 3: Evaluational
7) Testing and validation
8) Refinement and reclarification
9) Applying reassessment
GenAI Knowledge Co-construction
Accuracy
No
Yes
Human-GenAI Decision

Fig. 1. A step-by-step protocol for knowledge co-construction using GenAI tools.

- ***Dimension 1 - (Contextual):*** In the constructivist theory, knowledge construction relies on a contextual-centric interpretation of facts or ideas. Similarly, consideration of contextual factors is crucial in human-GenAI prompting as well. While crafting an effective prompt, the netizen's question, his/her background knowledge, and adaptive curiosity can help form initial ideas that facilitate the quality of their learning experiences by converting an inquiry to new concepts via the GenAI tools [4]. In the conceptual dimension, the first step is to construct a question as a prompting task, ranging from a simple inquiry to a more complex instruction. In addition, the second step is applied to integrate prior background knowledge into the drafted question by embedding keywords [30]. Then, the third step is deployed as the desire to improve the objectives of the question by changing initial words to meet her/his expectations.

- ***Dimension 2 - (Structural):*** An effective prompt should be a well-designed inquiry with a clear structure that supports knowledge construction between netizens and GenAI tools. In other words, a well-structured prompt facilitates the interpretation of the user interaction by improving cognitive processes while fostering human-GenAI symbiosis to co-construct knowledge. As a result of such symbiotic communication, the interactions between netizens and GenAI tools are enhanced, progressively augmenting their mutual engagements. In practice, the structure of a prompt concerns the detailed writing format of words and connected sentences, which requires following a proper layout. Some prompt structures have been introduced in the literature, such as chain-of-thought, zero-shot, prompting with instances (e.g., one- and multi-shot), and role prompting [4], i.e., netizens must learn when to apply these methods and which structure is the most proper for the expected outcome. In practice, netizens must define a well-structured prompt according to their expectations so that the GenAI tool can craft more relevant content that aligns with the inquiry while considering the above five key elements. Also, the next dimension of prompting protocol is the structural category in which the fourth step entails writing an initial prompt (e.g., a couple of sentences) according to the question; the fifth and sixth ones are considered to optimize and produce the most relevant and precise version of the prompt to be tested GenAI tools [30].

- ***Dimension 3 - (Evaluational):*** Evaluating the drafted prompt in the former dimension requires three more challenging steps (7-9) to validate, refine, and reassess it to gain the desired outcome [30]. During the testing and validation step, the user must verify and ensure the reliable accuracy of the GenAI-crafted content according to her/his expected objectives in the prompt, considering possible biases [3]. Furthermore, in the refinement and reclarification step, the user is involved in fine-tuning the structure of the prompt by increasing the clarity of the objectives and elaborating on complex expectations. When (re)assessing a GenAI-crafted content based on a given prompt, the user is not only required to recognize its suitability as the best match for her/his expectations, but (s)he also must analyze the possibility of biases may be caused by human-GenAI interactions [37] when pondering its accuracy and validity [21]. To validate the reliability of GenAI-crafted content, netizens can write a prompt to obtain its rationale by asking six questions to address those biases (see Figure 2). However, in some sensitive topics (e.g., scientific tasks [19]), the trustworthiness of GenAI results can not be confirmed unless relevant experts review them and approve their accuracy [33]. For example, in a recent study by [33], Ninety-one dichotomous questions (yes/no) were designed and classified into three difficulty classifications. Firstly, researchers randomly selected 20 questions from each class. Secondly, they applied the ChatGPT to find the answers to these sixty questions. Thirdly, they collected the responses from two endodontic experts who answered

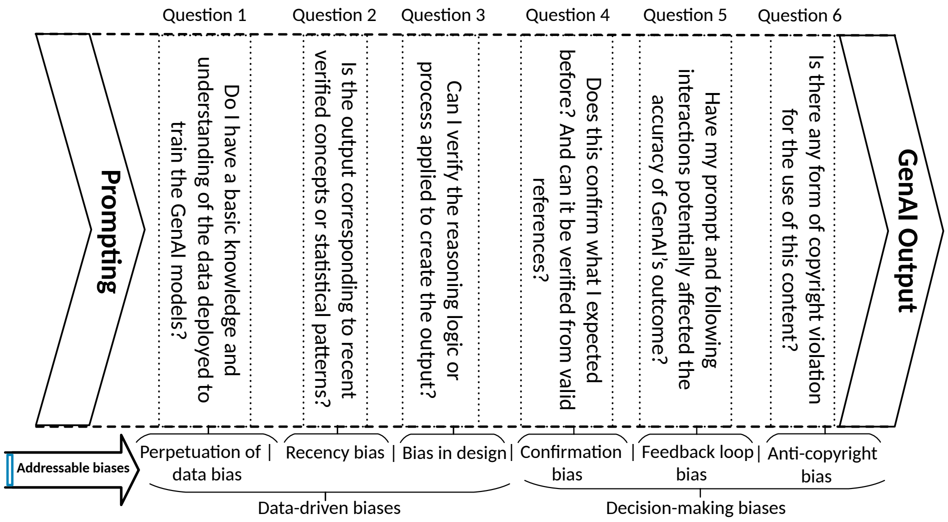
Question 1
Question 2
Question 3
Question 4
Question 5
Question 6
Prompting
Do I have a basic knowledge and understanding of the data deployed to train the GenAI models?
Is the output corresponding to recent verified concepts or statistical patterns?
Can I verify the reasoning logic or process applied to create the output?
Does this confirm what I expected before? And can it be verified from valid references?
Have my prompt and following interactions potentially affected the accuracy of GenAI's outcome?
Is there any form of copyright violation for the use of this content?
GenAI Output
Addressable biases
Perpetuation of data bias
Recency bias
Bias in design
Confirmation bias
Feedback loop bias
Anti-copyright bias
Data-driven biases
Decision-making biases

Fig. 2. Six-step question-based framework for validating the accuracy of GenAI output.

those 60 questions separately. Fourthly, they conducted a statistical analysis via the SPSS tool to evaluate the consistency and accuracy of the experts' answers compared to replies generated using the ChatGPT. Finally, they concluded that ChatGPT has achieved an average accuracy of 57.33%. This experiment highlights the fact that netizens must validate the accuracy of GenAI outcomes when using them for specialized applications without being overseen by relevant experts.

TABLE 3. A sample of a well-designed prompt using the three-dimensional prompting protocol.

Steps	Prompting Processes	Summary of ChatGPT-4o Output	Validation Question
Dimension 1 (Conceptual) [1, 2, and 3]	How can I implement a quantum-resistant TLS protocol in a Java-based Android application?	It generated a list of guidelines and source codes that suggest two libraries, such as **OQS-OpenSSL** and **Bouncy Castle**.	Which library is the best option for implementing the quantum-resistant TLS 1.3?
Dimension 2 (Structural) [4, 5, and 6]	What Java library can be used as an efficient way to implement a quantum-resistant TLS protocol in an Android application? Please give me optimized sample source codes to implement it.	It crafted a list of instructions and source codes on how to implement the quantum-resistant TLS protocol using the Bouncy Castle library.	Why the Bouncy Castle Library is the most efficient one?
Dimension 3 (Evaluational) [7, 8, and 9]	Why can the Bouncy Castle library efficiently implement a quantum-resistant TLS 1.3 protocol using Java in an Android app? Please give me optimized source codes to implement it and provide me with some references from developer.android.com to validate why it is efficient.	It provided five reasons for why the **Bouncy Castle** Library is an efficient way to implement a quantum-resistant TLS 1.3 protocol in an Android application. Also, it refers to three references that support the efficiency factors from Google's official source (developer.android.com) for Android application development.	Have my prompt and subsequent interactions effectively influenced the ChatGPT's output towards a better knowledge co-construction accuracy?

[*] Note that validation questions vary for different prompting tasks. These questions must clear possible biases in the GenAI-crafted content, especially when the output involves sensitive scientific or industrial concepts.

As depicted in Table 3, we constructed a well-designed and -structured prompt by deploying the prompting protocol to find and validate the efficient way of implementing the quantum-resistant TLS 1.3 in a Java-based Android application. In our experiments, despite forming a straightforward question in dimension 1, the ChatGPT-4o crafted two contents, where the correct one was the second, considering the experts' point of view. By following the instructions of the next two dimensions, we optimized and validated the results of ChatGPT-4o while revising the prompt two more times to reach valid or unbiased content. Note that the validation questions are customized based on the question-based framework according to expectations from the GenAI tool. After that, we recruited $n = 10$ volunteers (e.g., DevOps engineers) to test the sample in Table 3 by following the prompting protocol's dimensions. Eventually, all volunteers pointed out that they learned not to trust the first output of ChatGPT-4o and verify it by optimizing the prompt and validating the results (e.g., source codes). Although our test focused on a specific topic, it can be extended by creating more generalized samples in CWE programs to simplify the understanding of the prompting protocol for all netizens. Technically, the GenAI tools still have negative cyberpsychological impacts (e.g., distortion of netizens' critical thinking and digital trust) and technological limitations such as hallucinations, content integrity, privacy, safety, copyrights, and ownership that are yet to be addressed [39]. On the other hand, as depicted in Table 2, the current CWE programs partially cover the knowledge and skills necessary for all ranges of netizens to mitigate emerging risks (e.g., cryptocurrency heists and deepfake-driven social engineering attacks) in society. Therefore, educational institutes must upgrade the existing CWE programs by integrating more relevant creative thinking knowledge and skills (e.g., the above-suggested prompting protocol) and defense mechanisms (e.g., Take IT Down[3]) to help netizens mitigate emerging cyber risks.

5 Adaptive E-Governance

Understanding the fact why netizens trust deepfakes, which may lead them to be trapped in unprecedented cyberattacks, is a complex agenda [23]. The current literature on public engagement has emphasized that people refuse to accept scientific evidence when it risks their profits or questions their beliefs [28]. Over the last few years, multiple socio-economic damages have been caused by proof-of-work blockchains (e.g., Bitcoin) that negatively impact the climate crisis, human mortality, and financial losses (e.g., crypto heists via deepfake-phishing attacks [20]) in society. For instance, according to the Chainalysis Crypto-Crime report [6], more than $24.2 billion in value was received from illicit cryptocurrency addresses in 2023. Nevertheless, there was a significant drop in value compared to 2022 with $39.6 billion, highlighting that criminal activities were declined, even though the dark web markets and ransomware attacks have increased dramatically. Therefore, behaviors of unprecedented cyberthreats must be investigated and integrated into CWE programs. This, in turn, involves developing, upgrading, and teaching defense mechanisms and safe ways of accountable human-GenAI interactions (e.g., the above-suggested prompting protocol) at all levels of society: individuals, families, and schools. Therefore, enhancing the efficiency and effectiveness of CWE programs requires taking more proactive and strategic actions from educational systems and proactive attempts from the societal, industrial, and governmental sectors [17]. Figure 3 depicts a circular-based puzzle, which directs the following six proactive players to collaborate on upgrading the effectiveness of CWE programs in our society. In this holistic usable cybersecurity management framework, six groups of actors play a significant role by sitting at a decision-making table, instigating the problem, and contributing to the primary goal – mitigation of emerging cyber risks, where players' actions impact neighbors' decisions (in)directly and their following activities inherently.

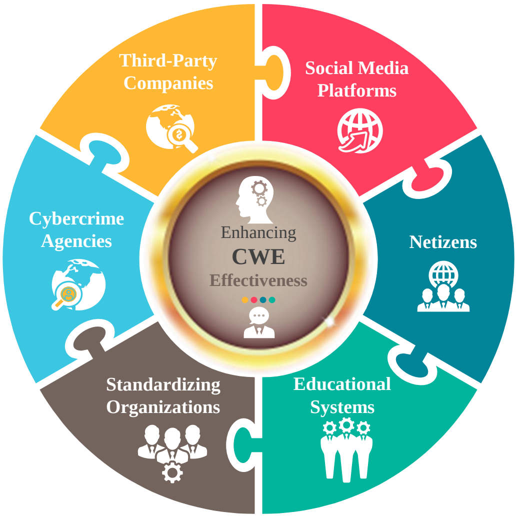

Fig. 3. Circular puzzle-based framework for enhancing the effectiveness of CWE programs.

(1) Third-party companies provide innovative GenAI-based services by typically considering only partially their impacts on the public, most likely because their priority is consumers' attraction to increase revenue. For example, GenAI tools are easily accessible to everyone, enabling netizens to construct content without considering ethical and social consequences [24] that require regulatory compliance and control by standardizing organizations. On the other hand, these companies are responsible for deploying explainable GenAI models and ensuring fairness while processing netizens' data according to the AI laws. Hence, they must update and incorporate the e-governance policies to reform their services for shaping safer human-GenAI interactions.

(2) Social media platforms are owned by private companies (e.g., Meta and ByteDance), which offer free social networking services to billions of netizens around the world, storing and processing consumers' sensitive data and activities [28]. Companies' usage of netizens' data must comply with international laws such as EU General Data Protection Regulation (GDPR) [11] and China Data Security Law (DSL) [10] if they want to continue their businesses in those regions. This highlights the fact that they must support research and development to reduce the potential risks of their services and prevent them from being fined under the laws. Also, they should develop and integrate deepfake detection algorithms in their platforms to enlighten netizens' awareness of fabricated media.

(3) Netizens are Internet users who surf online platforms where intruders continuously target them via cyberattacks. They should be necessarily subjected to CWE programs and trained, as it is their social responsibility to learn possible mechanisms and tools to deal with unprecedented risks [17]. Learning defense mechanisms and critical thinking skills can help netizens increase their awareness of deepfakes and their socio-economic impact and actively report them to relevant e-governance agencies. Similarly, parents must regularly learn defense mechanisms by participating in CWE programs and policies proactively and guide their young netizens to make the right choices and practice safer web-surfing activities.

(4) Cybercrime Agencies are executive governmental organizations (e.g., Europol's EC3 or EU EDPB) that are responsible for investigating cybercrimes, identifying the unprecedented cyberspace risks and controlling them according to regional/international laws. In addition, they must collect netizens' reports on their experiences with privacy violations and cooperate with educational systems to develop CWE programs based on recently discovered cyberthreats.

(5) Standardizing organizations are responsible for developing enforcement policies in all technical and nontechnical fields and creating uniformity across corresponding agencies, producers, and consumers. They also publish standard guidelines to be executed by target parties as responsible sectors in society. Moreover, governmental agencies have statutory duties to enforce adaptive e-governance policies and standards proactively by investigating public usage of GenAI tools and controlling the accountability of their risky services. Additionally, they must regularly adapt and introduce CWE policies to be integrated and carried out by educational systems or mass media.

(6) Educational systems are the executive and research centers for organizing CWE programs and training teachers and potential netizens. Also, they are responsible for gradually developing effective resources and practices to mitigate possible cyberspace risks by following the CWE policies. In some cases, they can propose adaptive policies for standardizing organizations to incorporate unprecedented risks of emerging GenAI-based technologies.

6 Robust Oversight

Below, we suggest three proactive recommendations for concerned policymakers as critical requirements in today's society that could best contribute to mitigating emerging cyber risks.

- AI legal policies and actions: Currently, several countries have already regulated AI policies and laws (e.g., controversial regulations in China that came into force in January 2023) or have been discussed (e.g., EU AI Act[9] [12] and USA's National AI Initiative Act of 2020[10]). Note that apart from national or region-based legal actions, it is necessary to have global policies and laws to

[9] https://www.europarl.europa.eu/

[10] https://www.ai.gov/

control AI-based technologies that impact all netizens worldwide. However, in practice, netizens are still concerned about how risky AI services are and what tools and guidelines are necessary to control such risks. Hence, CWE programs must be continuously developed and taught to all levels of society [17] by considering the socio-economic threats of emerging technologies, including GenAI services and deepfakes. Constitutionally, e-governance agencies must regularly monitor and control GenAI tools to identify potential risks that may endanger society and reform them by enforcing educational and preventive policies.

- **Free of charge up-to-date training and reliable tools:** It is of utmost importance to build awareness among netizens and other parties via cost-free training programs that would have a worldwide reach as well as to offer preventive tools and services that would effectively help them to distinguish between deepfakes and legitimate contents. It is advised that involved decision-makers or actors from different organizations (i.e., societal, industrial, and governmental) should work toward the formation of an international body where cybersecurity experts and policymakers can work together toward continuous research and development for providing up-to-date CWE programs and defense tools considering the six actors' decisions and actions (see Figure 3). One natural candidate for such a body is the Internet Society[11], a global nonprofit organization that aims to keep the Internet open, globally-connected, secure, and trustworthy.

- **Compliance with legal policies:** Enforcing and ensuring continuous compliance of GenAI tools with regulations also requires regular monitoring actions from governmental organizations to assure fairness and accuracy. In a sense, fairness and explainability can be interpreted as an AI governance, risk, and compliance (GRC) problem, in which regulatory agencies must periodically investigate the public usage of GenAI tools, their associated risks, and the accountability of their services. In practice, the service providers must prove to the regulatory organizations that their tools comply with policies by providing unbiased evidence and continuously adapting to integrate new GRC policies. For example, the recent rule proposed by the USA FCC [9] on September 9, 2024, aims to protect netizens from the abuse of deepfakes in the form of unwanted robocalls and robotexts, which could be deployed as preventive actions for controlling artificial ads and their impacts on the elections. In another regulatory action, legislatures across the USA[12] are passing urgently necessary laws to regulate deepfakes in electioneering communications, i.e., thirteen states have already enacted legislation, and 32 states have put forth bills.

7 Conclusion and Outlook

This article has focused on investigating the global-scale role of CWE (or digital media literacy) in defense to help netizens better understand and prepare for the threats of deepfakes and their associated risks. Regarding the development of up-to-date CWE programs and the legislation of enforcement policies and actions, policymakers and educational institutions appear to be a step back from malicious actors. To overcome such an issue, regulatory agencies and educational organizations should aim to: (1) investigate copyright and ownership issues of deepfakes more broadly to enhance defense mechanisms for netizens viewing misinformation, disinformation, and malinformation; (2) integrate recent policies, laws, and actions (e.g., EU AI Act[6]) into the CWE curriculum regularly, and upgrade the existing programs (e.g., including the suggested prompting protocol in Section 4) and create educational outlets for all range of netizens; (3) support the training of CWE curriculum in society more than ever and monitor regular integration of emerging socio-economic threats of deepfakes as a crucial part of such programs; (4) reduce the time between legislation of enforcement policies and proactive actions to be taken in any involved administrative

[11] https://www.internetsociety.org

[12] https://www.citizen.org/

or educational organizations; and (5) engineer practical solutions or tools that provide public services to recognize deepfakes and their reliability for everyone inquiring about such contents.

8 Acknowledgement(s)

This study was partially supported by the Ulam research grant No. BPN/ULM/2022/1/00069, funded by the Polish National Agency for Academic Exchange (Narodowa AgencjaWymiany Akademickiej (NAWA), and by U.S. NSF awards #1820609, #2114824, and #2131144.

References

[1] Altuncu, E., Franqueira, V. N., and Li, S. (2024). Deepfake: definitions, performance metrics and standards, datasets, and a meta-review. *Frontiers in Big Data*, 7:1400024.

[2] Angelis, L. D., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., and Rizzo, C. (2023). Chatgpt and the rise of large language models: the new ai-driven infodemic threat in public health. *Frontiers in Public Health*, 11:1166120.

[3] Bai, S., Gonda, D. E., and Hew, K. F. (2024). Write-curate-verify: A case study of leveraging generative ai for scenario writing in scenario-based learning. *IEEE Transactions on Learning Technologies*.

[4] Bulat, A. and Georgios, T. (2024). Language-aware soft prompting: Text-to-text optimization for few-and zero-shot adaptation of v &l models. *International Journal of Computer Vision*, 132(4):1108–1125.

[5] Burnett, T. (01/03/2023). Cambridge university among elite universities to ban chatgpt due to plagiarism fears.

[6] Chainanlysis (15/05/2024). The 2024 crypto crime report.

[7] Chen, H. and Magramo, K. (04/02/2024). The 2024 crypto crime report.

[8] Chen, X., Liu, T., Fournier-Viger, P., Zhang, B., Long, G., and Zhang, Q. (2024). A fine-grained self-adapting prompt learning approach for few-shot learning with pre-trained language models. *Knowledge-Based Systems*, page 111968.

[9] Commission, F. C. (08/02/2024). Implications of artificial intelligence technologies on protecting consumers from unwanted robocalls and robotexts.

[10] Creemers, R. (2022). China's emerging data protection framework. *Journal of Cybersecurity*, 8(1):tyac011.

[11] EDBP (15/05/2024). Dpb resolves dispute on transfers by meta and creates task force on chat gpt.

[12] Hutson, M. (2023). Rules to keep ai in check: nations carve different paths for tech regulation. *Nature*, 620(7973):260–263.

[13] Hyder, S. (15/05/2024a). The future workforce: How conversational ai is changing the game (liveperson).

[14] Hyder, S. (15/05/2024b). How are educators reacting to chat gpt?

[15] imothy R. McIntosh, Liu, T., Susnjak, T., Watters, P., Ng, A., and Halgamuge, M. N. (2023). A culturally sensitive test to evaluate nuanced gpt hallucination. *IEEE Transactions on Artificial Intelligence*.

[16] IProovT (15/05/2024). How to protect against deepfakes – statistics and solutions.

[17] Lewin, C., Niederhauser, D., Johnson, Q., Saito, T., Sakamoto, A., and Sherman, R. (2021). Safe and responsible internet use in a connected world: Promoting cyber-wellness. *Canadian Journal of Learning and Technology*, 47(4):n4.

[18] Lin, Z. (2024a). How to write effective prompts for large language models. *Nature Human Behaviour*, 8(4):611–615.

[19] Lin, Z. (2024b). Techniques for supercharging academic writing with generative ai. *Nature Biomedical Engineering*, pages 1–6.

[20] Lindburg, S. (15/05/2024). Quantum ai review, fake quantum ai scam by elon musk exposed!

[21] Linehan, M., Byers, C., Brooks, N. N., and Freeman, L. (2024). Responsible generative ai. *Industrial Internet Consortium*.

[22] Lo, L. S. (2023). The clear path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship*, 49(4):102720.

[23] Mazurczyk, W., Lee, D., and Vlachos, A. (2024). Disinformation 2.0 in the age of ai: A cybersecurity perspective. *Communications of the ACM*, 67(3):36–39.

[24] Menekse, M. (2023). Envisioning the future of learning and teaching engineering in the artificial intelligence era: Opportunities and challenges. *Journal of Engineering Education*, 112(3):578–582.

[25] Metze, K., Morandin-Reis, R. C., Lorand-Metze, I., and Florindo, J. B. (2024). Bibliographic research with chatgpt may be misleading: the problem of hallucination. *Journal of Pediatric Surgery*, 59(1):158.

[26] NewYorkPost (13/03/2024). Will deepfake ai content influence the 2024 election?

[27] Office, F. F. (15/05/2024). Malicious actors manipulating photos and videos to create explicit content and sextortion schemes.

[28] Osborne, J. and Pimentel, D. (2022). Science, misinformation, and the role of education. *Science*, 378(6617):246–248.

[29] Puig, A. (15/05/2024). Scammers use ai to enhance their family emergency schemes.

[30] Robertson, J., Ferreira, C., Botha, E., and Oosthuizen, K. (2024). Game changers: A generative ai prompt protocol to enhance human-ai knowledge co-construction. *Business Horizons*.

[31] Sayantan, B., Logan, N. S., Davies, L. N., Sheppard, A. L., and Wolffsohn, J. S. (2023). Assessing the utility of chatgpt as an artificial intelligence-based large language model for information to answer questions on myopia. *Ophthalmic and Physiological Optics*, 43(6):1562–1570.

[32] Security, H. (24/08/2021). Media literacy and critical thinking online.

[33] Suárez, A., Díaz-Flores, G. V., Algar, J., Sánchez, M. G., de Pedro, M. L., and Freire, Y. (2024). Unveiling the chatgpt phenomenon: Evaluating the consistency and accuracy of endodontic question answers. *International endodontic journal*, 57(1):108–113.

[34] Suh, M., Youngblom, E., Terry, M., and Cai, C. J. (2021). Ai as social glue: uncovering the roles of deep generative ai during social music composition. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–11.

[35] Umair, M., Bouguettaya, A., Lakhdari, A., Ouzzani, M., and Liu, Y. (2024). Exif2vec: A framework to ascertain untrustworthy crowdsourced images using metadata. *ACM Transactions on the Web*, 18(3):1–27.

[36] Wang, M., Wang, M., Xu, X., Yang, L., Cai, D., and Yin, M. (2023). Unleashing chatgpt's power: A case study on optimizing information retrieval in flipped classrooms via prompt engineering. *IEEE Transactions on Learning Technologies*.

[37] Xu, M., Du, H., Niyato, D., Kang, J., Xiong, Z., Mao, S., Han, Z., Jamalipour, A., Kim, D. I., Shen, X., et al. (2024). Unleashing the power of edge-cloud generative ai in mobile networks: A survey of aigc services. *IEEE Communications Surveys & Tutorials*.

[38] Yang, X. and Zhang, M. (2024). Genai distortion: The effect of genai fluency and positive affect. *arXiv preprint arXiv:2404.17822*.

[39] Zhang, P. and Boulos, M. N. K. (2023). Generative ai in medicine and healthcare: Promises, opportunities and challenges. *Future Internet*, 15(9):286.

[40] Zhou, M., Abhishek, V., and Srinivasan, K. (01/03/2023). Bias in generative ai (work in progress).

9 Authors titles

Milad Taleby Ahvanooey is an Assistant Professor and Ulam scientist at Warsaw University of Technology, Poland. Prior to this, he was a senior researcher at Nanyang Technological University, Singapore.

Wojciech Mazurczyk is a University Professor at Warsaw University of Technology, Warsaw, Poland.

Dongwon Lee is a Professor at The Pennsylvania State University, University Park, PA, USA

Received 9 July 2024; revised 26 October 2024; accepted 22 January 2025