DOI:10.1145/3715317

To mitigate the risks of deepfakes, enhanced cyber-wellness programs are needed that empower both producers and consumers of generative AI–based content.

BY MILAD TALEBY AHVANOOEY, WOJCIECH MAZURCZYK, AND DONGWON LEE

Socioeconomic Threats of Deepfakes and the Role of Cyber-Wellness Education in Defense

DUE TO THE limits of science and its steep learning curve, we must rely on the expertise of others to develop our knowledge and skills.[26] Toward this end, social media platforms have revolutionized how netizens—users who are actively engaged in online communities—gain knowledge and skills by facilitating the exchange of costless information with the public (for example, followers or influencers). Businesses around the world also use these platforms along with tools based on generative artificial intelligence (GenAI) to craft synthetic media, hoping to grow revenue by attracting more customers and improving their online experience.[28]

These GenAI tools can also be leveraged by malicious actors, who create deepfakes—content that has been digitally manipulated to convincingly imitate people or things, often to portray events that have not actually happened. Such synthetic media may also contain partial facts to mislead viewers, often called *misinformation*. But public awareness of these threats is lagging. In a 2022 survey conducted by iProov[15] of consumers (n = 16,000) from eight countries, only 71% knew what a deepfake video is and only 57% could recognize the difference between a real and an artificially crafted video. According to a poll report published by the *New York Post*,[31] a survey of 2,000 registered U.S. voters showed not only that participants are increasingly pessimistic about the use of deepfakes in political advertising, but also that they could not distinguish between AI-generated and human-created content.

To highlight the danger of mass-crafted spurious information, in June 2023 a group of two engineers used open source GenAI models to devel-

» key insights

- **Generative AI tools can empower cyber threats and have cyberpsychological effects on netizens, allowing malicious actors to craft deepfakes in the form of disinformation, misinformation, and malinformation.**
- **Service providers not only must enhance GenAI tools to reduce hallucinations, but they also have a statutory duty to mitigate data-driven biases.**
- **To counteract the socioeconomic threats of deepfakes, we must enhance the effectiveness of cyber-wellness education programs; these programs should teach practices for human-GenAI knowledge co-construction and content validation to reduce decision-making biases.**

IMAGE BY ANDRIJ BORYS ASSOCIATES WITH ASSISTANCE OF SHUTTERSTOCK.AI


DEEPFAKES

op *CounterCloud*,[a] an experimental "propaganda machine." In an exemplary test case, the project's GenAI-crafted deepfake texts[1,32] (for example, 50 tweets and 20 news articles daily) were seen as convincing 90% of the time. The primary goal behind this project was to enlighten people about how effortless it is to weaponize GenAI to propagate deliberately manipulated content, called *disinformation*, on a global scale. In addition, the creators claimed that by spending approximately $4,000 per month, anyone can generate more than 200 articles per day. This is enough content to counter the efforts of more than 40 news outlet—without requiring any human interaction—and could be enough to influence an election.

In addition to disinformation and misinformation, there is also *malinformation*, forms of deepfake threats[33] based on manipulated facts intended to harm targeted victims (for example, cyberscamming[19] and cyberbullying[34]). Netizens must therefore be alert to these suspicious activities, whereby malicious actors can use malinformation to target them through vulnerabilities in their decision making processes (for example, cognitive biases, which can lead to poor judgment).

Socioeconomic Threats of Deepfakes

As shown in recent cases reported by mass media and law enforcement agencies,[34] scammers can deceive victims by creating and deploying deepfake content (for example, voice and/or video) that poses as a colleague, family member, or a celebrity known to the target group or individual. When netizens lack the necessary awareness, knowledge, and skills to identify deepfakes, they may eagerly trust and be fooled by this synthetic content, leading to unwise actions and/or financial losses. For example, in a recent cyberscamming incident using GenAI tools, $25.6 million was stolen from a multinational firm in Hong Kong by tricking an employee into believing that the CEO requested the funds transfer via a video call.[8] Similarly, if a victim clicks on the link connected to a deepfake ad and transfers some money or cryptocurrency, they may fall into a scammer's phishing trap.[19] Recently, in a consumer alert issued by the U.S. Federal Trade Commission (FTC),[27] 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 voice recording of their family member—often available on a social media profile—and then apply a GenAI tool (for example, the LivePerson voice AI chatbot[14]) to create a deepfake voice recording that is virtually indistinguishable from the real one. To avoid these voice-cloning traps, potential victims should not trust these messages and should confirm their authenticity[34] by apply cyber-wellness knowledge, such as calling the claimed person and asking for verification.

Regulatory agencies and industry must jointly develop educational programs to guide people in making safe decisions when dealing with deepfake content.

In another form of cyber scam, crafted deepfake videos of celebrities were widely spread on social media platforms to promote services linked to phishing websites (for example, the Quantum AI Elon Musk trading bot[19]). Moreover, in a June 2023 public alert,[34] the U.S. Federal Bureau of Investigation (FBI) announced that they had received reports on cybercriminals deploying GenAI tools to create artificial porn videos and photos of underage victims. The offenders sent the generated content directly to victims for cyberbullying or sextortion purposes. To combat these unprecedented threats, the U.S. National Center for Missing and Exploited Children has built an online platform called Take It Down,[b] which provides free services to help victims stop and prevent the spread of these videos and images, particularly sexually explicit content with underage children.

These cyber threats are an ongoing worldwide phenomenon with significant implications for many areas, including crypto market trading, elections, health, and education. In a recent regulatory action,[35] the U.S. Federal Communications Commission (FCC) declared that unwanted GenAI-crafted robocalls and robotexts should be banned by law. It also closed a comment period on a public citizen's petition for a new rule concerning the application of deepfakes in election

a https://www.youtube.com/watch?v=cwGdkrc9i2Y

b https://takeitdown.ncmec.org/

advertisements. This legislative action was initiated after the propagation of a robocall on January 21, 2024, in which GenAI-created audio impersonating President Joe Biden discouraged citizens from voting. It is expected that once this petition is approved at the national level, such regulations will be recommended internationally under the aegises of organizations such as the European Union and the United Nations, as well as tech-related consortiums like the Christchurch Call.[c]

Educational Aspects

While the full dangers of these technologies are beyond anyone's imagination, GenAI tools also have some legitimate uses. Netizens can easily access GenAI software for free or pay a subscription fee for various applications (see Table 1). These GenAI tools provide opportunities to innovate and reform industries, such as education and advertising. At the same time, they can negatively affect individuals' lifelong learning and can have severe consequences for their critical and creative thinking skills. A recent study showed that OpenAI's GPT-4 could achieve high scores on standardized tests, such as 99% on the GRE Verbal and 89% on the SAT Math,[24] due to it applying more collaborative, creative, and efficient transformers compared to previous versions. In addition, a survey study showed that more than 51% of students believed applying GenAI software such as ChatGPT to pass exams or prepare multimedia assignments is technically cheating.[12]

Public usage of such unaccountable and unexplainable[28] GenAI tools has given rise to societal and governmental concerns over why they are easily accessible to everyone. For instance, several countries recently banned ChatGPT in public universities by blocking access to it through their networks. They also ruled that using GenAI tools to prepare assignments would be considered academic misconduct. This includes at least five Australian states and eight elite universities in the Russell Group in the U.K., including Oxford and Cambridge.[6] But since GenAI tools are rapidly evolving, crafted outcomes are becoming impossible to distinguish from actual human-generated media. Consequently, regulatory agencies and industry must jointly develop educational programs to guide people in making safe decisions when dealing with deepfake content. In an exemplary policy action, Italy's Data Protection Agency temporarily restricted ChatGPT's services. 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 suspicious media.

To address this issue, the science of *cyber-wellness education (CWE)* uses standardized guidelines to teach netizens how to protect themselves and stay safe while interacting in cyberspace.[16] Over the past few decades, the European Union[d] and many individual countries (for example, the U.S.[e] and the U.K.[f]) have developed digital literacy and/or CWE programs (see Table 2), actively using them in educational institutions to reduce the risk of cyber threats. Numerous online scams recently caused netizens to lose millions of dollars,[19] mainly due to various forms of deepfake cyber threats. The current CWE programs are not sufficiently updated to protect against these threats, making them inadequate for netizens, even if they do receive CWE training. Moreover, these programs often focus on training students and pay less attention to regular consumers of digital content on the Internet.

As outlined in Table 1, GenAI tools and social media platforms allow people to easily craft and spread propaganda and manipulate information, often using intentionally crafted

c https://www.christchurchcall.org/

d https://www.digital-wellbeing.eu

e https://lincs.ed.gov/

f https://www.gov.uk/government/publications/online-media-literacy-strategy

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 content based on a prompt.	ChatGPT, Brad, and ChatSonic	+ Creative, business, and academic writing + Source-code generation for programming + Textual content translation + Real-time 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 - Cyberscamming, cyberbullying, and sextortion - Copyright and ownership violation of copyrighted images
MLLMs can compose or edit songs (for example, 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 music - Misinformation, disinformation, and malinformation
LMMs can craft or edit videos or any other multimedia content 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 model generation + Game design and development - Cyberscamming, cyberbullying, and 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) that netizens must input to guide the GenAI tool in constructing or editing content, whether it is word-based text, image, audio, or a combination of these media.

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

Action Plans	Details	List of Contents
Digital-Wellbeing Education[d] (Co-founded by the Erasmus+ program of the EU)	This curriculum consists of CWE course materials, which are aimed at 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 well-being.	- 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[e] (Founded by the U.S. Department of Education)	The U.S.'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[f](Founded by the U.K. Department for Science, Innovation, and Technology, and the Department for Digital, Culture, Media & Sport)	This program aims to train and empower netizens across the U.K. to control their safety through online platforms. More than 170 organizations are currently involved in delivering CWE in the U.K. This strategy outlines the government's plan to coordinate the media literacy landscape in several years and provides a CWE framework for best practices for evaluating the content and delivery of up-to-date educational materials.	- Data and Privacy - Online Environment - Information Consumption - Online Consequences - Online Engagement

deepfakes to deceive their target audiences. In addition to these intentional deepfakes, traditional media outlets, online libraries, and social media platforms can be both perpetrators and victims of unintentionally shared fabricated content. Sometimes users of GenAI tools will write prompts, overtrust the result, and unintentionally use it for sensitive purposes.[33] This is not a safe practice, as all output from GenAI tools should be validated due to the possible biases and hallucinations[28] of the underpinning models. For instance, a New York attorney used ChatGPT to conduct legal research to represent a client's injury claim. While overseeing the suit, the federal judge noticed that six citations quoted in the attorney's brief were falsified.[g] In this hallucination scenario, ChatGPT made up the cited cases and even asserted that they were accessible in major legal databases.

GenAI Distortion Risks in Society

GenAI tools' ability to craft seemingly realistic media by combining fact with fiction may distort netizens' perceptions of the content's reality and its associated dangers—a phenomenon known as *GenAI distortion risks*.[38] One way to reduce these risks is to learn how to apply prompting knowledge and skills correctly, as well as understand decision-making biases that help people design and refine effective prompts for crafting more accurate outcomes from GenAI tools and avoiding deepfakes.[36] Technically, this involves crafting or refining a query: one or more connected sentences given to GenAI tools that may result in producing more accurate and relevant content. The validity and quality of output from GenAI tools are influenced by two independent factors, the lack of which can negatively affect accuracy:[17]

- *Coherence:* the quality of a prompt being logically close to what a user expects to gain as content.
- *Relevance:* the availability of relevant data that can be processed and collected by Web-scraping approaches through GenAI tools.

Enhancing netizens' knowledge and skills on how to craft and refine prompts can lead to more accurate results and, eventually, better content being produced by large language models (LLMs), large vision language models (LVLMs), music-specialized large language models (MLLMs), speech large language models (SLLMs), and large multimodal models (LMMs). Since GenAI tools deploy Web-scraping techniques to collect relevant data from resources accessible on the Internet, they are somewhat limited. Therefore, if the prompts fed to the GenAI models contain insufficient or incorrect data, they may produce biased content containing irrelevant concepts. Technically, these biased outcomes are caused by three *data-driven biases:*[37]

- Problematic training data unintentionally perpetuates biases that may link inaccurate or incorrect data to specific subjects.
- Accessibility to real-time data is a significant limitation, which can cause recency bias.
- Underpinning models can be biased by design and then embed biases into the associated data, which inevitably produces unfair outcomes.[39]

In addition to these data-driven biases, there are three *decision-making biases* to which netizens are vulnerable:

- The tendency to over-trust AI tools, which leads to a false confirmation.
- Optimizing prompts, which might steer GenAI models to adapt their re-

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

plies toward netizens' objectives and can cause feedback-loop bias.[28]

▸ Using GenAI-crafted content for commercial purposes (for example, journalism or scientific publications), which might violate the ownership rights of similar content and can be interpreted as anti-copyright bias.

Practically, data-driven biases may cause GenAI hallucinations that result in unintentional deepfakes if netizens trust this content while also being subject to decision-making biases.

To reduce the possibility of creating invalid content using GenAI tools, we will now discuss a prompting protocol that can guide netizens to understand the validity and reliability of GenAI-crafted information. Crafting content requires well-structured human-GenAI interactions to operate accurately and achieve high-standard results. Therefore, to write an effective prompt, the following five elements must be considered:

▸ *Relevance:* An effective prompt must be relevant to the expected task by providing adequate information to direct the GenAI tool to make precise predictions.[21]

▸ *Diversity:* An effective prompt should contain a range of specific details to ensure the generalization of the integrated GenAI models to new data.[9]

▸ *Consistency:* An effective prompt should follow a correct, consistent format to guide the GenAI models to successfully perform the requested tasks.[9]

▸ *Simplicity:* An effective prompt should be concise and coherent to reduce the possibility of confusion for GenAI models when processing the requested tasks.[28]

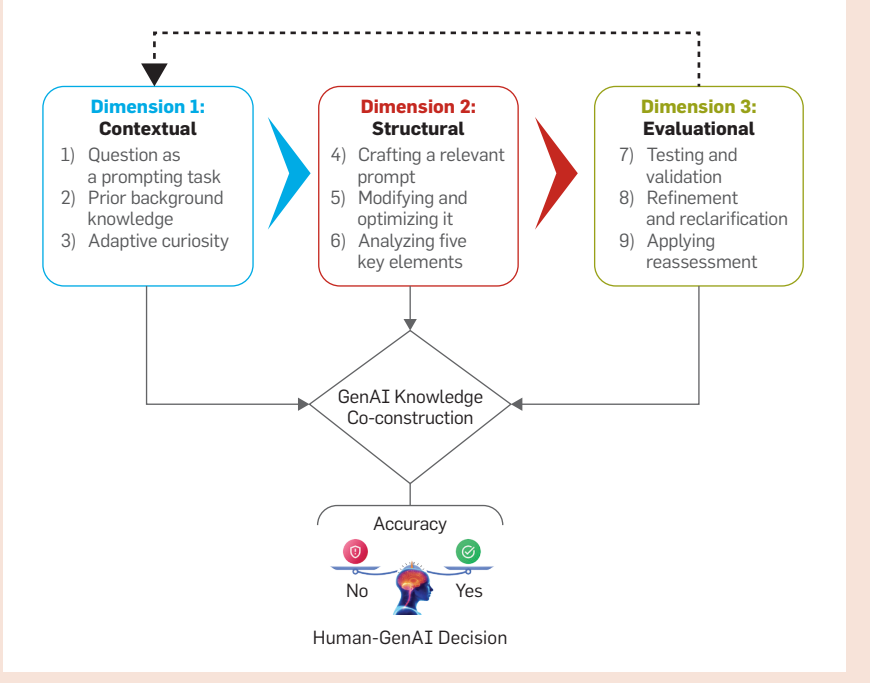
▸ *Clarity:* An effective prompt should be clearly and unambiguously defined so that the task concept meets the best possible quality of transparency that the GenAI model can perform.

Effective prompts can result in more informative and accurate GenAI outcomes, while defectively designed ones may result in irrelevant and confusing responses.[4] 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 these tools without forfeiting their integrity and ingenuity. But since GenAI tools technically generate content by interpreting a given prompt and reproducing new information based on the trained data, there is a possibility they may construct incorrect or misleading results (hallucinations).[25] This is a common problem with GenAI tools, as they are developed to craft new forms of content and can sometimes create plausible-sounding content that might actually be incorrect.[23] 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. Since the capability of GenAI tools depends on interpreting netizens' prompts, they should be treated as collaborative entities in a knowledge co-construction process.[30] To enrich human-GenAI knowledge co-construction, we extended the prompting protocol introduced by Robertson et al.[28] It helps netizens to write more detailed and effective prompts and includes a question-based framework to validate their accuracy.

In constructivist theory, learners construct knowledge instead of passively receiving information. Commonly, they will 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.[28] In the context of human-GenAI prompting, an iterative refinement process is necessary to improve the dialogue between GenAI tools and netizens, formulating a cohesive reply to an inquiry as a problematic task. Considering the above constructivist perspectives,[28] the prompting process can therefore be formed as an iterative, 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). In constructivist theory, knowledge construction relies on a context-centric interpretation of facts or ideas. Similarly, consideration of contextual factors is also crucial in human-GenAI prompting. While crafting an effective prompt, the user's question, background knowledge, and adaptive curiosity can help form the initial ideas that facilitate the quality of their learning experience by converting an inquiry to new concepts via GenAI tools.[5] In the conceptual dimension, the first step is to construct a question as a prompting task, ranging

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



from a simple inquiry to a more complex instruction. The second step is applied to integrate prior background knowledge into the drafted question by embedding keywords.[28] The third step improves the objectives of the question by changing initial words to meet the user's 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. This symbiotic communication enhances the interactions between netizens and GenAI tools, progressively augmenting their mutual engagement. In practice, the structure of a prompt involves the detailed writing of words and coherent sentences, which requires following a proper format. Some prompt structures have been introduced in the literature, such as chain-of-thought, zero-shot, prompting with instances (for example, one-shot and multi-shot), and role prompting.[5] Netizens must learn when to apply these methods and which structure is best suited 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 five key elements listed above. Within this structural dimension, the fourth step entails writing an initial prompt (for example, a couple of sentences) according to the question; the fifth and sixth steps help to optimize and produce the most relevant and precise version of the prompt to be tested with GenAI tools.[28]

Dimension 3 (evaluational). Evaluating the prompt drafted in the previous dimension requires three more challenging steps (7–9) to validate, refine, and reassess it to gain the desired outcome.[28] During the testing and validation step, the user must verify and ensure the reliable accuracy of the GenAI-crafted content according to their expected objectives in the prompt, considering possible biases.[3] 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 GenAI-crafted content based on a given prompt for accuracy and validity,[20] the user must not only determine whether it is the best match for their expectations, but also consider the possibility of biases caused by human-GenAI interactions.[37] To validate the content's reliability, netizens can obtain its rationale by asking six questions to address those biases (see Figure 2). However, with some sensitive topics (for example, scientific tasks[18]), the trustworthiness of GenAI results cannot be confirmed unless relevant experts review them.[29] For example, in a recent study,[29] 91 dichotomous (yes/no) questions about endodontics were designed and divided into three difficulty classifications. First, researchers randomly selected 20 questions from each class. Then they asked ChatGPT for answers to those 60 questions, while also getting responses from two endodontic experts, who answered the 60 questions separately. Finally, they conducted a statistical analysis via the SPSS tool to evaluate the consistency and accuracy of the experts' answers compared with the replies generated using ChatGPT. They concluded that ChatGPT had an average accuracy of 57.33% on the selected questions. This experiment highlights the fact that netizens must validate the accuracy of GenAI outcomes when using them for specialized applications while not being overseen by relevant experts.

As depicted in Table 3, we used the prompting protocol to construct a well-designed and well-structured prompt to find and validate an efficient way of implementing quantum-resistant TLS 1.3 in a Java-based Android application. In our experiments, despite posing a straightforward question in dimension 1, ChatGPT-4o created two responses, of which the second one, which considered the experts' point of view, was correct. By following the steps of the next two dimensions, we optimized and validated the results of ChatGPT-4o while revising the prompt two more times to obtain valid and 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 (for example, DevOps engineers) to test the sample in Table 3 by following the prompting protocol's three dimensions. Eventually, all volunteers pointed out that they learned not to trust the first output of ChatGPT-4o and to verify it by optimizing the prompt and validating the results (for example, source code). 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 (for example, distortion of netizens' critical thinking and digital trust) and technological limitations, such as hallucinations, as well as content

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

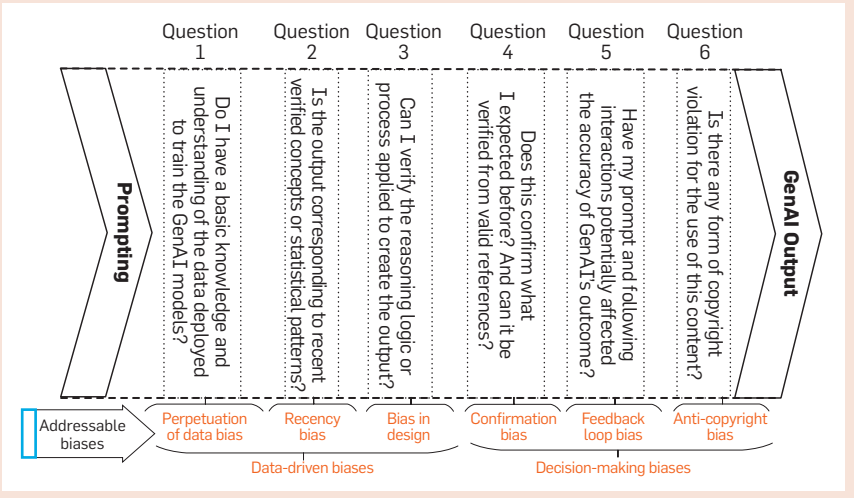


Table 3. A sample of a well-designed prompt using the three-dimensional prompting protocol.

Steps	Prompting Processes	Summary of ChatGPT-4o Output	Validation Question
Dimension 1 (Conceptual; steps 1,2,3)	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; steps 4,5,6)	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 is the Bouncy Castle Library the most efficient one?
Dimension 3 (Evaluational; steps 7,8,9)	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 ChatGPT's output toward 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.

integrity, privacy, safety, copyright, and ownership issues that have yet to be addressed.[40] At the same time, as depicted in Table 2, current CWE programs partially cover the knowledge and skills necessary for all types of netizens to mitigate the emerging risks in society (for example, cryptocurrency heists and deepfake-driven social engineering attacks). Therefore, educational institutions must upgrade existing CWE programs by integrating more relevant creative-thinking knowledge and skills (such as the above-suggested prompting protocol) and defense mechanisms (such as Take It Down) to help netizens address emerging cybersecurity risks.

Adaptive E-Governance

Understanding why netizens trust deepfakes is a complex issue.[22] The current literature on public engagement emphasizes that people refuse to accept scientific evidence when it risks their profits or questions their beliefs.[26] Over the past few years, much damage has been caused by proof-of-work blockchains (for example, Bitcoin) that negatively affect the climate, human mortality, and personal finances (for example, crypto heists via deepfake phishing attacks[19]). For instance, according to the Chainalysis Crypto-Crime report,[7] more than $24.2 billion was received from illicit cryptocurrency addresses in 2023. This was a significant drop compared with 2022, during which $39.6 billion was transferred, highlighting that criminal activities have declined despite dark web markets and ransomware attacks increasing dramatically. Nevertheless, these numbers provide further evidence that these unprecedented cyber threats must be investigated and integrated into CWE programs. This, in turn, involves developing, upgrading, and teaching defense mechanisms and safe, accountable ways of conducting human-GenAI interactions (for example, the suggested prompting protocol outlined earlier) at all levels of society: individuals, families, and schools. Enhancing the efficiency and effectiveness of CWE programs therefore requires more proactive and strategic action from educational institutions as well as society, industry, and government.[16]

Figure 3 depicts a circular-based puzzle representing collaboration on upgrading the effectiveness of CWE programs among various groups. In this holistic, usable cybersecurity management framework, six groups of actors play a significant role by sitting at a decision-making table, investigating the problem, and contributing to the primary goal: mitigation of emerging cybersecurity risks, where players' actions (in)directly affect their neighbors' decisions and subsequent activities. The roles of each of these six groups are discussed below.

Third-party companies provide innovative GenAI-based services typically by only partially considering their impacts on the public, most likely because their priority is attracting consumers to increase revenue. For example, GenAI tools are easily accessible to everyone, enabling netizens to construct content without considering ethical and social consequences,[24] which require regulation and control by standards organizations. At the same time, these companies are responsible for deploying explainable GenAI models and ensuring fairness while processing netizens' data according to AI laws. As such, they must update and incorporate e-governance policies to reform their services toward shaping safer human-GenAI interactions.

Social media platforms are owned by private companies like Meta and ByteDance, which offer free social networking services to billions of netizens around the world while storing and processing consumers' sensitive data and activities.[26] Companies' usage of netizens' data must comply with international laws, such as the EU's General Data Protection Regulation (GDPR)[11] and China's Data Security Law (DSL)[10] if they want to continue to do business in those regions. This highlights the need for research and development to reduce the potential risks of their services and prevent them from being fined under the laws. They should also develop and integrate deepfake detection algorithms in their platforms to enhance netizens' awareness of GenAI-fabricated media.

Netizens are Internet users who surf and engage through online communities, where intruders target them via cyberattacks. They should be exposed to CWE programs and trained accordingly, as it is their social responsibility to learn possible mechanisms and tools to deal with today's cybersecurity risks.[16] Learning defense mechanisms and critical thinking skills from these programs can help netizens increase their awareness of deepfakes and their

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

socioeconomic impacts as well as actively report them to relevant e-governance agencies. Similarly, parents must regularly learn these techniques by participating in CWE programs and proactively guide their children to make the right choices and practice safer Web-surfing activities.

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

Standards 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 standards guidelines for individuals and organizations to follow as responsible societal actors. Governmental agencies have statutory duties to enforce adaptive e-governance policies and standards proactively by investigating the public usage of GenAI tools and controlling the accountability of their risky services. They must also regularly adapt and introduce CWE policies to be integrated and carried out by educational systems or through mass media campaigns.

Cybercrime agencies are executive governmental organizations, such Europol's EC3 or the EU's EDPB, that are responsible for investigating cybercrimes, identifying the latest cyberspace risks, and addressing them with 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 cyber threats.

Robust Oversight

Here, we present three proactive recommendations for concerned policymakers that could best contribute to mitigating emerging cybersecurity risks.

AI legal policies and actions. Currently, several countries have already regulated AI policies and laws (for example, controversial regulations in China that came into effect in January 2023) or have been discussed (for example, the EU's AI Act[h,13] and the U.S.'s National AI Initiative Act of 2020). Note that in addition to national or region-based legal actions, it is necessary to also have global policies and laws to control AI-based technologies that affect all global netizens. However, despite these legal efforts, netizens are still concerned about the risks of these AI technologies and what tools and guidelines are needed to address them. Hence, CWE programs must be continuously developed and taught at all levels of society[16] by considering the socioeconomic 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 these tools 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, via cost-free training programs with worldwide reach, about the risks of deepfakes as well as to offer preventive tools and services that help them effectively distinguish between deepfakes and legitimate content. Decision-makers and actors from different organizations (that is, 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 that consider the six actors' decisions and actions (see Figure 3). One natural candidate for such a body is the Internet Society,[i] a global nonprofit organization that aims to keep the Internet open, globally connected, secure, and trustworthy.

Compliance with legal policies.

h https://www.europarl.europa.eu/portal/en
i https://www.internetsociety.org

Enforcing and ensuring the compliance of GenAI tools with regulations also requires regular monitoring actions from governmental organizations to ensure 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, service providers must prove to regulatory organizations that their tools comply with policies by providing unbiased evidence and continuously adapting to integrate new GRC policies. For example, a recent rule proposed by the U.S. FCC[35] 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 potential impact on elections. In another regulatory action, legislatures across the U.S.[j] are urgently passing necessary laws to regulate deepfakes in electioneering communications; 13 states have already enacted legislation, and 32 states have put forth bills.

Conclusion and Outlook

In this article, we investigated the global-scale role of CWE (or digital media literacy) to help netizens better understand and prepare for the threats of deepfakes and their associated risks. Unfortunately, when it comes to 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 behind malicious actors. To overcome this problem, regulatory agencies and educational organizations should take a number of actions. First, they should investigate the copyright and ownership issues of deepfakes more broadly to enhance defense mechanisms for netizens viewing misinformation, disinformation, and malinformation. They should also integrate recent policies, laws, and actions (for example, the EU's AI Act) into the CWE curriculum regularly, upgrade existing programs (for example, including the prompting protocol suggested earlier), and create educational outlets for all types of netizens. Further, as a crucial part of these programs, they should expand the training of CWE curricula and monitor the regular integration of emerging socioeconomic threats of deepfakes. Also important is reducing the time between the legislation of enforcement policies and proactive actions that can be taken by any involved administrative or educational organizations. Finally, these organizations must engineer practical solutions that provide public services to recognize deepfakes and ensure the reliability of GenAI-generated content for anyone inquiring about such content.

j https://www.citizen.org/

Acknowledgments

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. Deepfake: Definitions, performance metrics and standards, datasets, and a meta-review. *Frontiers in Big Data 7* (2024), 1400024.
2. Angelis, L.D. et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health 11* (2023), 1166120.
3. Bai, S., Gonda, D.E., and Hew, K F. Write-curate-verify: A case study of leveraging generative ai for scenario writing in scenario-based learning. *IEEE Transactions on Learning Technologies 17* (2024), 1313–1324.
4. Biswas, S. et al. Assessing the utility of ChatGPTas an artificial intelligence-based large language model for information to answer questions on myopia. *Ophthalmic and Physiological Optics 43*, 6 (2023), 1562–1570.
5. Bulat, A. and Georgios, T. Language-aware soft prompting: Text-to-text optimization for few- and zero-shot adaptation of v & l models. *Intern. J. of Computer Vision 132*, 4 (2024), 1108–1125.
6. Burnett, T. Cambridge University among elite universities to ban ChatGPT due to plagiarism fears. *Cambridge News* (Mar. 1, 2023).
7. Chainanlysis. The 2024 crypto crime report (2024).
8. Chen, H. and Magramo, K. The 2024 crypto crime report (2024).
9. Chen, X. et al. A fine-grained self-adapting prompt learning approach for few-shot learning with pre-trained language models. *Knowledge-Based Systems 299* (2024), 111968.
10. Creemers, R. China's emerging data protection framework. *J. of Cybersecurity 8*, 1 (2022), tyac011.
11. EDBP. EDPB resolves dispute on transfers by Meta and creates task force on chat GPT (Apr. 13, 2023).
12. Gordon, C. How are educators reacting to ChatGPT? *Forbes* (Apr. 30, 2023).
13. Hutson, M. Rules to keep AI in check: nations carve different paths for tech regulation. *Nature 620*, 7973 (2023), 260–263.
14. Hyder, S. The future workforce: How conversational AI is changing the game (liveperson). *Forbes* (May 17, 2021).
15. IProov. Statistics and solutions: How to protect against deepfakes (2023).
16. Lewin, C. et al. Safe and responsible internet use in a connected world: Promoting cyber-wellness. *Canadian J. of Learning and Technology 47*, 4 (2021).
17. Lin, Z. How to write effective prompts for large language models. *Nature Human Behaviour 8* (2024), 611–615.
18. Lin, Z. Techniques for supercharging academic writing with generative AI. *Nature Biomedical Engineering 9* (2024), 426–431.
19. Lindburg, S. Quantum AI review, fake quantum AI scam by Elon Musk exposed! (2024).
20. Linehan, M. et al. Responsible generative AI. *Industrial Internet Consortium* (2024).
21. Lo, L.S. The clear path: A framework for enhancing information literacy through prompt engineering. *The J. of Academic Librarianship 49*, 4 (2023), 102720.
22. Mazurczyk, W., Lee, D., and Vlachos, A. Disinformation 2.0 in the age of AI: A cybersecurity perspective. *Communications of the ACM 67*, 3 (2024), 36–39.
23. McIntosh, T.R. et al. A culturally sensitive test to evaluate nuanced GPT hallucination. *IEEE Transactions on Artificial Intelligence 5*, 6 (2023).
24. Menekse, M. Envisioning the future of learning and teaching engineering in the artificial intelligence era: Opportunities and challenges. *J. of Engineering Education 112*, 3 (2023), 578–582.
25. Metze, K. et al. Bibliographic research with ChatGPTmay be misleading: The problem of hallucination. *J. of Pediatric Surgery 59*, 1 (2024), 158.
26. Osborne, J. and Pimentel, D. Science, misinformation, and the role of education. *Science 378*, 6617 (2022), 246–248.
27. Puig, A. Scammers use AI to enhance their family emergency schemes. Federal Trade Commission (Mar. 20, 2023).
28. Robertson, J. et al. Game changers: A generative AI prompt protocol to enhance human-AI knowledge co-construction. *Business Horizons 67*, 5 (2024).
29. Suárez, A. et al. Unveiling the ChatGPT phenomenon: Evaluating the consistency and accuracy of endodontic question answers. *Intern. Endodontic J. 57*, 1 (2024), 108–113.
30. Suh, M. et al. AI as social glue: Uncovering the roles of deep generative AI during social music composition. In *Proceedings of the 2021 CHI Conf. on Human Factors in Computing Systems*. ACM (2021), 1–11.
31. SWNS. Will deepfake AI content influence the 2024 election? *New York Post* (Mar. 13, 2024).
32. Umair, M. et al. Exif2vec: A framework to ascertain untrustworthy crowdsourced images using metadata. *ACM Transactions on the Web 18*, 3 (2024), 1–27.
33. U.S. Department of Homeland Security. Media literacy and critical thinking online (2021).
34. U.S. Federal Bureau of Investigation. Malicious actors manipulating photos and videos to create explicit content and sextortion schemes (Jun. 5, 2023).
35. U.S. Federal Communications Commission. Implications of artificial intelligence technologies on protecting consumers from unwanted robocalls and robotexts. Federal Register (2024).
36. Wang, M. et al. Unleashing ChatGPT's power: A case study on optimizing information retrieval in flipped classrooms via prompt engineering. *IEEE Transactions on Learning Technologies 17* (2024).
37. Xu, M. et al. Unleashing the power of edge-cloud generative AI in mobile networks: A survey of AIGC services. *IEEE Communications Surveys & Tutorials 26*, 2 (2024).
38. Yang, X. and Zhang, M. GenAI distortion: The effect of GenAI fluency and positive affect. arXiv preprint arXiv:2404.17822 (2024).
39. Zhou, M., et al. Bias in generative AI (work in progress). arXiv preprint arXiv:2403.02726 (2023).
40. Zhang, P. and Boulos, M.N.K. Generative AI in medicine and healthcare: Promises, opportunities and challenges. *Future Internet 15*, 9 (2023), 286.

Milad Taleby Ahvanooey (m.taleby@ieee.org) 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.