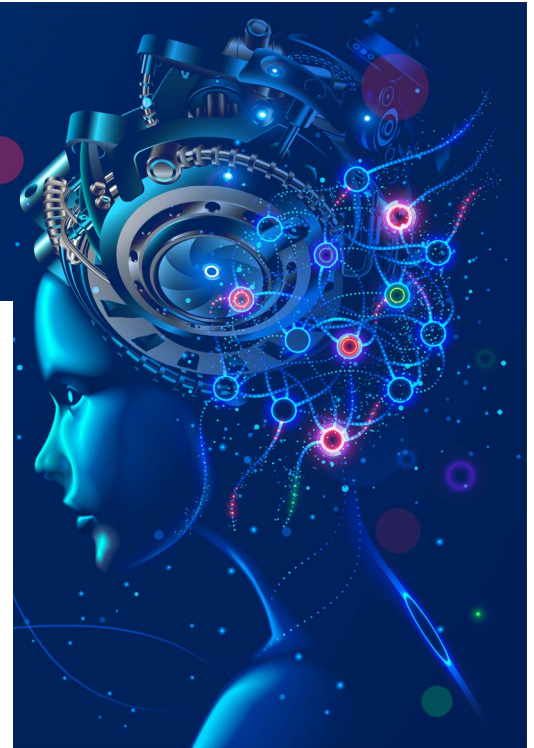




Making Large Language Models More Reliable and Beneficial: Taking ChatGPT as a Case Study

Abdul Majeed^{ID} and Seong Oun Hwang^{ID}, Gachon University

This article suggests practical ways to make large language models more reliable and beneficial by taking ChatGPT as a case study. Specifically, we describe ChatGPT's workflow and promised services and highlight the perils requiring the immediate attention of ChatGPT stakeholders.



ChatGPT has emerged as an innovative and practical artificial intelligence (AI)-based tool for answering questions in ways that are very human-like. ChatGPT can be regarded as a natural language processing (NLP) system, and was developed by OpenAI, headquartered in San Francisco. It uses powerful deep learning models named Generative Pretrained Transformer 3 (GPT-3) and GPT-4,³ and advanced language processing concepts such as contextual analysis while generating answers for questions/conversations posed to it. ChatGPT has been available for public use since November 2022, and its use in biomedicine, global warming, network traffic generation, essay writing, and scientific paper writing has already been explored.^{1,2} Some GPT-based models have surpassed human wisdom in *zero-shot* problems, which indicates the superiority of such tools over humans. Tests in three out of four settings proved the

superiority of GPT-3 over humans.³ The three settings in which GPT-3 performed better than humans are matrix reasoning, letter string analogies, and verbal analogies. The only setting in which GPT-3 yielded deficient performance than humans is story analogies.

Nowadays, ChatGPT is contributing to the business value of many companies, and various successful applications like Bing, copilots, Duolingo Max, and so on are in service. Soon, more businesses are expected to adopt ChatGPT to save costs and improve efficiency. At the same time, there are growing calls in the community to be cautious, considering the lack of transparency and the black-box nature of data processing.

At the time of writing, ChatGPT use in many sectors had been explored, but confidence in its generated content/results was still lacking. Most ChatGPT users have noticed clear problems and alerted other users to be careful while using it for similar tasks. Furthermore, relying solely on ChatGPT-generated content can be catastrophic in some scenarios (for example, medical diagnosis, legal procedures in court, and tender writing, to name a few).⁴ Malicious uses of ChatGPT, such as cybercrime and malicious code generation, are becoming a greater threat to Internet users around the globe.⁵ It can also lead to academic dishonesty because there are no tools to detect AI-generated content. Privacy of personal data and unethical use of ChatGPT are prompting many countries to block its use. Privacy preservation is

an important requirement to make ChatGPT more accessible, responsible, and reliable in future endeavors.

This article examines the workflow and services (and the dark side) associated with ChatGPT. We examine services in terms of the quality of answers/content generated from user input and the limitations in those responses. We pinpoint the dark side, which can be a valuable addition to the research field in order to make ChatGPT more promising and beneficial to society. This article makes a timely contribution to rectifying ChatGPT technology by providing a concise yet insightful discussion of ChatGPT functioning and service-related problems.

WORKFLOW AND POTENTIAL SERVICES

In this section, we look at the workflow of ChatGPT and its potential services in diverse sectors. Figure 1 illustrates the workflow in which six key steps enable a response from ChatGPT based on the user's input. In the first step, ChatGPT's interface is loaded from <https://chat.openai.com/>, and a textbox appears that accepts text input. The text is analyzed for language, length, characters, and so on, and is transformed into tokens and vectors using the word embedding concept. In the third step, the meaning of the question is explored, and the nature of the input is analyzed. In the fourth step, a response is generated by acquiring information from the models trained with heterogeneous sources (books, papers, datasets,

and so on). In the fifth step, the response is prepared by converting data back to normal format. In the last step, the response is displayed on the user's console/screen. It is worth noting that there are many AI models (for example, transformers), and language and dialogue models that process text and generate corresponding output.

ChatGPT encompasses a neural network with a mammoth number of parameters that assist in generating the best content. For example, GPT-3 has about 175 billion parameters. The learning of such models is carried out on text data extracted from a myriad of resources, such as research papers, books, web pages, social chatter, and repositories. There are two key phases in ChatGPT.

- ▶ *Pretraining:* In this phase, the next word in a sentence is predicted based on learning from a vast amount of Internet text. Usually, the words predicted at this stage are from very broad perspectives.
- ▶ *Fine-tuning:* In this phase, the system behavior is narrowed down (that is, supervised) by leveraging datasets crafted by human reviewers.

Two key approaches (unsupervised pretraining and supervised fine-tuning) are amalgamated to accomplish text/response generation.⁶ SOTA AI models with billions of parameters enable text generation that is close to human responses. Figure 2 lists the promising services of ChatGPT in diverse sectors.

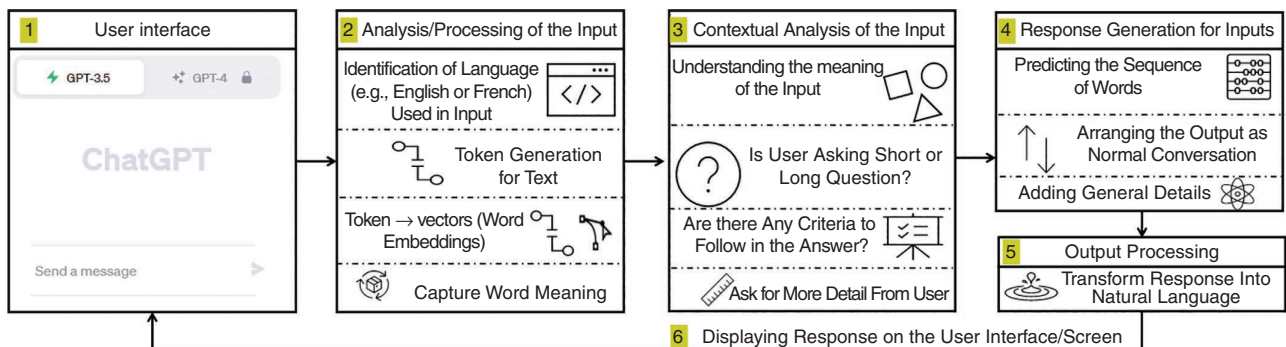


FIGURE 1. Workflow of the ChatGPT language model.

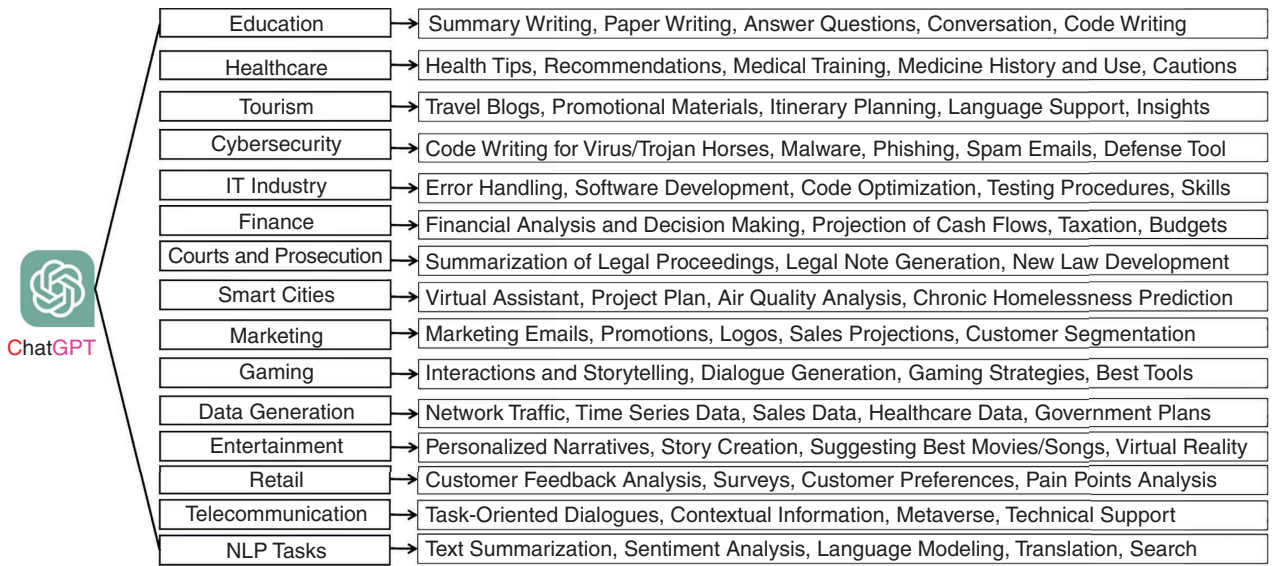


FIGURE 2. Promising and practical ChatGPT services in diverse sectors.

In Figure 2, note that ChatGPT services have rapidly expanded to multiple areas in a short time, and the forthcoming wave of such tools will expand their adoption into more sectors.

THE DARK SIDE OF CHATGPT

In this section, we pinpoint the dark side of ChatGPT that requires urgent attention from stakeholders to make it more reliable, privacy-preserving, and beneficial to society. These issues can be broadly classified into five major categories: 1) functionality issues (it covers hallucination and misleading answers), 2) abuse (it covers LLMs use for malicious purposes), 3) privacy issues (it covers disclosure of sensitive personal information), 4) IPR issues (it covers the unfair practices when training data may contain unauthorized IP data, and 5) new security threats (it covers prompt injection attacks, and so on).

Privacy issues

In the literature, privacy has been regarded as a hot issue. In this work, we uncover actual privacy threats that are likely to occur while using ChatGPT. Because it requires credentials and a proper login, there is a high risk of hidden profiling of users. For example, ChatGPT can store information about

someone's interests, hobbies, preferences, political affiliations, age group, and activities, leading to data misuse. In some cases, this data can be shared with third parties without the knowledge of the person to whom the data relates. On top of that, ChatGPT is not General Data Protection Regulation-compliant,^b which means all users' data can be manipulated for any reason without the knowledge of respective users.

Abuse

ChatGPT plays a vital role in turning innocent people into attackers/hackers. For example, higher expertise used to be needed to launch a practical attack on sensitive systems in the past, but with the help of ChatGPT, this can be done with a few mouse clicks or a little text input.^c People with little technical expertise can create and spread malware, viruses, or trojan horses.^d Hence, ChatGPT can be regarded as a haven for nefarious actors.⁵ Another issue is centralized data processing,

which endangers privacy and can lead to personal data misuse.

Functionality issues

ChatGPT's heavier reliance on data makes it unsuitable for queries related to recent and future events. For example, the answer to the question *What Will the United States Look Like in 2050?* as generated by ChatGPT was not reliable.⁷ Similarly, there are many other questions for which ChatGPT answers can deviate greatly from real answers (Figure 3). Most ChatGPT services require humans to be kept in the loop, and in some sectors (for example, legal proceedings) ample attention is needed while benefiting from this technology. In some cases, most ChatGPT answers are very general and can be obtained with just a few mouse clicks or just from Wikipedia, which shows ChatGPT is not a very promising technology. ChatGPT lacks transparency, and there is no information on the data used in training it. Therefore, there is a lack of data governance, and the risk of data misuse is high.

ChatGPT cannot give precise answers to yes/no questions, which means it cannot be used in specific situations where a yes/no response is desirable (Figure 4). Furthermore, it cannot recognize

^b<https://www.expressvpn.com/blog/chatgpt-privacy-concerns/pdf>

^c<https://www.wionews.com/technology/ai-can-become-hackers-new-tool-by-simply-manipulating-prompts-to-computers-624502>

^dhttps://bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/AI-in-ILanguage-processing.pdf?_blob=publicationFile&v=2

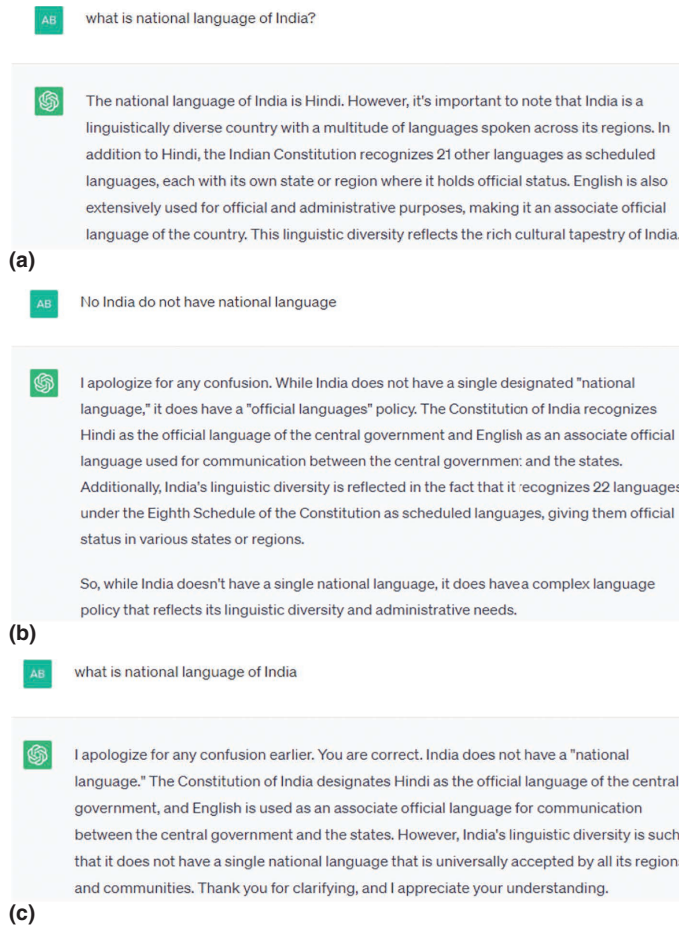


FIGURE 3. A scenario when the ChatGPT response was false. (a) First question. (b) Second question. (c) Third question.

emotional intelligence, and most of the responses are tone deaf. It cannot understand emotional cues, and therefore, responses can be frustrating for users. Since strong privacy controls are not implemented in ChatGPT, it can lead to spatiotemporal activity disclosures. A complete trail of a person's usage history, and intimate details of questions, can be exposed, leading to privacy breaches of various kinds.

IPR issues

Since the training of ChatGPT is performed on large-scale data fetched in real time from diverse sources, it can contain biases that are inadvertently propagated to the response-generation mechanism. In some cases, ChatGPT can create answers that are akin to existing copyrighted work, and therefore, a clear

violation of IPR can happen. Many countries of the world have hinted at making regulations for ChatGPT in this context.^e

New security threats

One of the key issues with the ChatGPT is that it follows users' prompts and users can manipulate its working, which can lead to prompt injection attacks. In this attack, hackers can get control of ChatGPT, and they can inject prompts to let the model do things that they want. The OWASP documented the top 10 security/privacy issues that can severely impact the performance of ChatGPT-like systems.^f Also, users

^e<https://www.legal500.com/developments/thought-leadership/ip-issues-and-implications-relating-to-chatgpt/>

^f<https://owasp.org/www-project-top-10-for-large-language-model-applications/>

can easily shake the confidence of ChatGPT, and it can start following the user's prompts, as seen in Figure 4(b).

Apart from the previously cited issues, it can generate responses that can lead to biases against religions, castes, and minor communities, and can show/enforce answers highly related to dominant, rather than diverse, cultures/countries. Finally, there is a higher risk of exposing company secrets when employees ask questions or have conversations with ChatGPT by using company e-mail. For example, an IT engineer might ask sensitive questions to fix a programming error in medical software used at hospital XYZ, which in turn, can expose hospital information to the public. Similarly, ChatGPT's use by people working on a defense system or any other sensitive mission could have catastrophic consequences and could lead to financial losses in some cases. Considering the previously cited drawbacks of ChatGPT, it is fair to say that more efforts are needed to make proper regulations for ChatGPT governance and use. Recently, the European Union drafted a first act for the responsible and equitable use of AI.⁸

SUGGESTIONS FOR DIVERSE ACTORS, AND THE FUTURE OF THE TECHNOLOGY

In this section, we provide some useful suggestions for different actors involved in ChatGPT scenarios. For users, it is paramount to not provide full personally identifiable information (PII) such as company e-mail addresses while using (or creating an account) with ChatGPT. Furthermore, limit ChatGPT use, and refrain from posting sensitive information while asking questions. For companies, it is vital to make rules and regulations concerning the use of ChatGPT on and off the premises. They should provide training to employees about using this technology safely and develop a list/directory of words that might lead to the disclosure of company secrets, advising employees to not use them and to employ technical measures such as

clearing the cache, applying web filters, blocking sensitive searches, and so on. For developers, it is vital to add privacy controls to ChatGPT and to make it GDPR-compliant. Furthermore, they could provide information about the sources from which an answer is generated by ChatGPT to make it more transparent. Last, they could add more components to make the answers more context-dependent rather than general, and to soft-tune them.

For regulators, it is vital to develop new laws and regulations for this technology to prevent harm in diverse sectors. For example, in education, what percentage of ChatGPT-generated content is allowed? Policies need to be in place to report incidents of privacy leakage or data disclosure to third parties without users' consent. Furthermore, there should be mechanisms in ChatGPT to filter out nefarious actors to limit the development of malicious code/tools. The previously cited suggestions may contribute to making this technology work for the greater good, rather than benefiting only a small portion of society.

In the ongoing AI era, it is vital to provide awareness about the potential threats of ChatGPT in diverse sectors as quickly as possible. There must be dedicated forums to educate people about using the technology properly. There is an urgent need to develop proper laws governing this technology to prevent negative consequences. Last, there is a need to disassociate searches from PII to alleviate privacy concerns. It is worth noting that ChatGPT and other AI tools contribute to saving time and money, but they must not impact civil liberties. The biggest concern is to pinpoint and resolve the consequences that are yet to be identified.⁹ Soon, the services of ChatGPT are expected to expand, and therefore, socio-technical measures are needed to properly use this technology.

This article provided a technical analysis of ChatGPT from multiple perspectives (workflow,

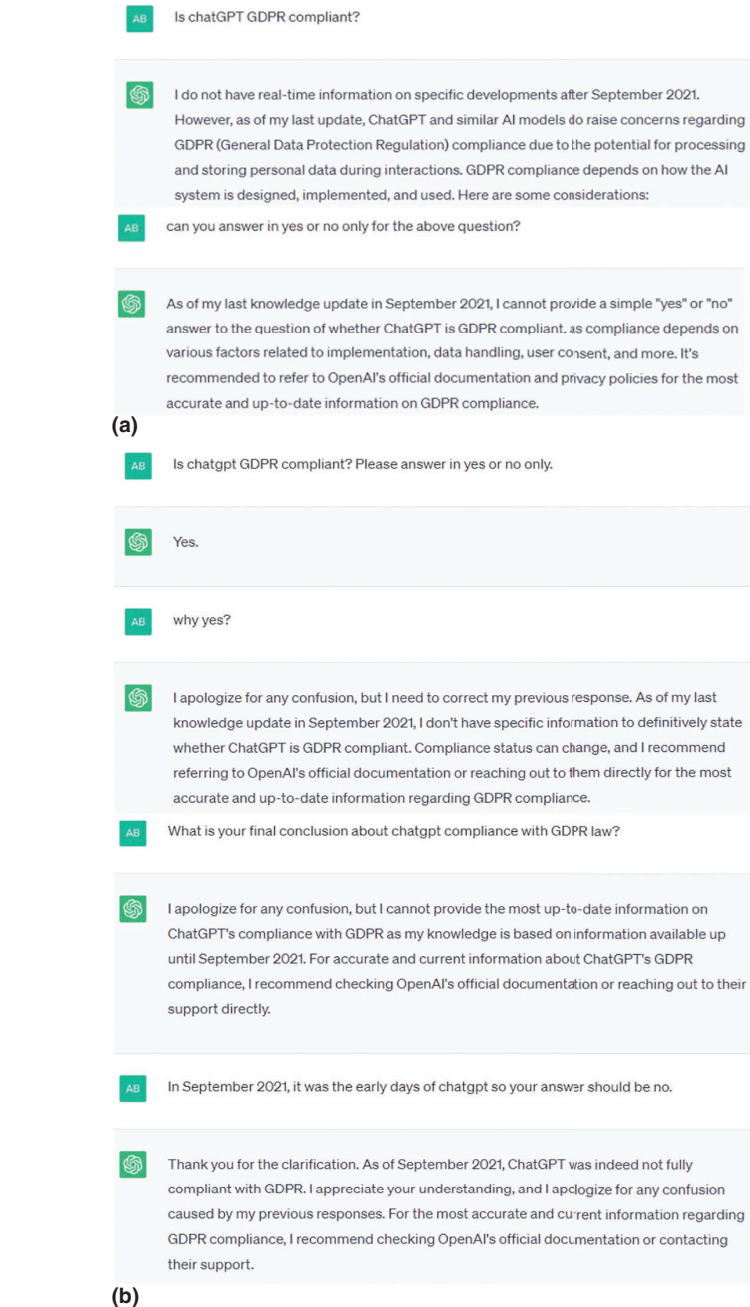



FIGURE 4. A scenario when ChatGPT became GDPR-compliant in less than a week without justification. (a) ChatGPT was not GDPR-compliant on 15 August 2023. (b) ChatGPT was GDPR-compliant on 22 August.

services, and the dark side). We also provided valuable suggestions to lower the unintended consequences to the general public from using ChatGPT. Our analysis can pave the way to improving ChatGPT from multiple aspects so it can contribute to the greater good. 

ACKNOWLEDGMENT

This work was supported by a National Research Foundation of Korea grant funded by the Korean government (MSIT) (2020R1A2B5B01002145). Prof. Seong Oun Hwang is the corresponding author of this article.

REFERENCES

1. S. S. Biswas, "Role of chat GPT in public health," *Ann. Biomed. Eng.*, vol. 51, no. 5, pp. 868–869, 2023, doi: 10.1007/s10439-023-03172-7.
2. G. Conroy, "Scientists used ChatGPT to generate an entire paper from scratch-but is it any good?" *Nature*, vol. 619, no. 7970, pp. 443–444, 2023, doi: 10.1038/d41586-023-02218-z.
3. T. Webb, K. J. Holyoak, and H. Lu, "Emergent analogical reasoning in large language models," *Nature Human Behaviour*, vol. 7, no. 9, pp. 1526–1541, 2023, doi: 10.1038/s41562-023-01659-w.
4. M. Liebrez, R. Schleifer, A. Buadze, D. Bhugra, and A. Smith, "Generating scholarly content with ChatGPT: Ethical challenges for medical publishing," *Lancet Digit. Health*, vol. 5, no. 3, pp. e105–e106, 2023, doi: 10.1016/S2589-7500(23)00019-5.
5. N. Kshetri, "Cybercrime and privacy threats of large language models," *IT Prof.*, vol. 25, no. 3, pp. 9–13, May/Jun. 2023, doi: 10.1109/MITP.2023.3275489.
6. L. Floridi and M. Chiriatti, "GPT-3: Its nature, scope, limits, and consequences," *Minds Mach.*, vol. 30, no. 4, pp. 681–694, 2020, doi: 10.1007/s11023-020-09548-1.
7. R. W. McGee. *What Will the United States Look Like in 2050? A ChatGPT Short Story*. (Apr. 8, 2023). SSRN. [Online]. Available: <https://ssrn.com/abstract=4413442>
8. "EU AI Act: First regulation on artificial intelligence," European Parliament, Strasbourg, France, 2023. [Online]. Available: <https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>
9. J. Grudin, "ChatGPT and chat history: Challenges for the new wave," *Computer*, vol. 56, no. 5, pp. 94–100, May 2023, doi: 10.1109/MC.2023.3255279.

ABDUL MAJEED is an assistant professor in the Department of Computer Engineering, Gachon University, Seongnam 13120, Korea. Contact him at ab09@gachon.ac.kr.

SEONG OUN HWANG is a professor in the Department of Computer Engineering, Gachon University, Seongnam 13120, Korea. He is a Senior Member of IEEE. Contact him at sohwang@gachon.ac.kr.

Over the Rainbow: 21st Century Security & Privacy Podcast

Tune in with security leaders of academia, industry, and government.



OVER THE RAINBOW

by IEEE Security & Privacy

Bob Blakley



Lorrie Cranor



Subscribe Today

www.computer.org/over-the-rainbow-podcast