Some discussions on critical information security issues in the artificial intelligence era

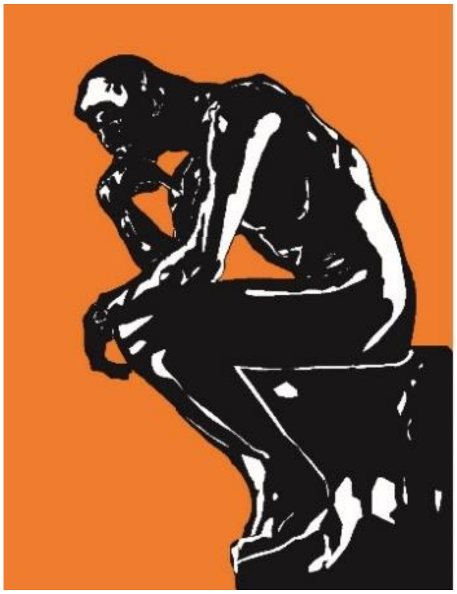
Vuong Quan Hoang [1], Viet-Phuong La [1,2], Hong-Son Nguyen [3], Minh-Hoang Nguyen [1,*]

[1] Centre for Interdisciplinary Social Research, Phenikaa University, Hanoi, Vietnam

[2] A.I. for Social Data Lab (AISDL), Vuong & Associates, Hanoi, Vietnam

[3] Office of CPV Central Committee, Hanoi, Vietnam

* Correspondence: hoang.nguyenminh@phenikaa-uni.edu.vn



March 30, 2024

[Original paper]

* * *

"Pressing the buttons has gradually become somewhat of a new technological ritual."

In "Innovation"; *The Kingfisher Story Collection* (2022)

Abstract

The rapid advancement of Information Technology (IT) platforms and programming languages has transformed the dynamics and development of human society. The cyberspace and associated utilities are expanding, leading to a gradual shift from real-world living to virtual life (also known as cyberspace or digital space). The expansion and development of Natural Language Processing (NLP) models and Large Language Models (LLMs) demonstrate human-like characteristics in reasoning, perception, attention, and creativity, helping humans overcome operational barriers. Alongside the immense potential of artificial intelligence (AI) are new security loopholes and more complex information security risks. As society is still trying to transition to a new phase to adapt to technological changes, the AI revolution continues to unfold, necessitating a reconsideration of the trajectory of societal transformation as it could exacerbate the aforementioned information security risks. Specifically, how should society evolve to keep pace with the transformative impact of the current AI technology wave? How can we manage and harness their power while ensuring information security as our presence in the virtual world increases? This article aims to shed light on and address these questions.

Keywords: artificial intelligence; societal transitions; AI limitations; human errors; power; freedom; regulation; Social Contract Theory; Mindsponge Theory

1. Technological advancements, security challenges, and information security

The rapid progress of Information Technology (IT) platforms and programming languages has transformed the dynamics and development of human society. Cyberspace and its accompanying utilities are increasingly expanding, leading to a gradual shift from life in the real world to life in the virtual world (also known as cyberspace or digital space). As of 2023, the Internet of Things (IoT) has connected approximately 15.14 billion devices globally (Vailshery, 2023). On average, each person on Earth now owns about 1.89 devices, nearly 24 times more than 20 years ago (with an average of 0.08 devices per person in 2003) (Lu & Da Xu, 2018). This number is predicted to nearly double by 2030, with about 29.42 billion devices connected. The variety of devices will become increasingly diverse, equipped with sensor systems or controllers to better interact with humans and integrated with artificial intelligence (AI) to assist in decision-making, searching, and transmitting information to users.

In the context where economic and social activities are increasingly well-connected through IoT, and soon with the potential integration of AI into almost every aspect of life in both the real and virtual worlds, not only individuals but also businesses and nations will face unprecedented challenges regarding information security risks (Keck et al., 2022).

As IT systems, especially the Internet, become integrated into life, a vast amount of information will be created, stored, and transmitted, such as personal information, social media interactions, business information, transaction data, insurance, health records, etc. Once this data is leaked, it can be exploited to defraud and negatively impact the lives of individuals, the operations of businesses, and the stability and sustainable development of nations. The World Economic Forum's (WEF) 2023 Global Risk Report ranked cybercrime and cybersecurity challenges among the top 10 risks currently and in the future (World Economic Forum, 2023). Indeed, cyber-attacks have increased by 600% since the start of the COVID-19 pandemic, with more than 5.4 billion malware attacks alone in 2022 (RiskXchange, 2023). According to Cybersecurity Ventures, cybercrime activities are expected to cause economic damage of approximately $10.5 trillion annually starting from 2025. These damages include data loss, stolen money, productivity decline, loss of intellectual property, theft of personal and financial data, fraud, disruptions following attacks on normal business processes, investigation after an attack, restoration and deletion of attacked data and systems, and reputation damage (Morgan, 2022). This is just a direct estimate of economic damage without considering the indirect impacts on the global economic and social system (Chính & Hoàng, 2009).

As society increasingly aims for the convenience and utility of smart cities, integrating a large number of electronic devices and software into life to manage assets, resources, and services becomes an inevitable trend. Thus, information will be collected from every citizen, device, building, and operational system to help monitor and manage traffic systems, power plants, water supply systems, waste treatment systems, information transmission systems, schools, hospitals, security, and other social services (Musa, 2018; Paiho et al., 2022). With the deep and complex linkage of device and software systems, cyber-attacks can quickly paralyze or partially paralyze the operation of society and nations or take control of the system if cybersecurity is not ensured. Hackers attacking utility management systems and taking control of devices, as occurred with the Uconnect system, a digital feature for entertainment, navigation, phone calls, and Wi-Fi access in vehicles, in 2015, using a security flaw to remotely control the vehicle to shut down or adjust its speed, compromising user safety. This led to several US car companies deciding to recall over a million vehicles in use, resulting in significant economic losses (Greenberg, 2015).

Furthermore, the types of information created in cyberspace can be regarded as a new kind of resource that facilitates the creation of extremely large and diverse datasets (big data). These datasets can be analyzed to identify connections, patterns, and trends in human behavior and social interactions. Currently, every minute, there are 6.3 million information searches on Google, over 527,000 photos shared on Snapchat, 456,000 tweets on the X platform (formerly Twitter), over 46,000 photos uploaded to Instagram, and approximately 510,000 comments posted and 293,000 new statuses updated on Facebook (Marr, 2021; Wise, 2023). Through the use of complex analytical techniques and algorithms, this massive amount of data can be utilized to reveal the thoughts, feelings, and behaviors of social media users, and employ this information for sophisticated psychological and behavioral manipulation schemes (Ho & Vuong, 2023). Moreover, as humans increasingly form emotional attachments to characters, assets, and applications within cyberspace, they can be more easily influenced in terms of psychology, emotions, and behavior (Mantello et al., 2023; Vuong et al., 2023a, 2023b).

A prime example of this is the case of Cambridge Analytica, a consulting firm that collected personal data from tens of millions of Facebook users and sold it to campaigns for emotionally manipulative political purposes, influencing the outcomes of elections. This scandal not only exposed the power held by those in control of information resources, especially technology corporations, but also how such power can be used to impact the operation of the economy, society, and politics (Liaropoulos, 2020; Nilekani, 2018).

Most recently, the launch of ChatGPT 3.5 on November 30, 2022, marked the beginning of what many experts are calling the “AI era.” Just one month after its debut, ChatGPT attracted over 100 million users, making it the fastest-growing software application in history (Hu, 2023). This user explosion has spurred the release of other competitive AI products, including Gemini, Ernie Bot, LLaMA, Claude, and Grok in 2023. In fact, AI technology has been widely applied in various aspects of life for some time now, such as in scientific research, healthcare, finance, entertainment, education, and transportation. Notable AI-powered applications we use almost daily include advanced web search tools (e.g., Google Search) and recommendation systems (used by YouTube, Amazon, and Netflix). However, operational capabilities (e.g., IT expertise requirements) and accessibility (e.g., high costs) remain significant barriers to societal understanding of AI as well as its functions in everyday life.

The expansion and development of Natural Language Processing (NLP) models and Large Language Models (LLMs) have showcased human-like features in reasoning, cognition, attention, and creativity, helping humans overcome operational barriers (Lappin, 2023; Vuong et al., 2023). Tasks that once required the operation of IT experts can now be completed by ordinary people through simple daily language commands. Additionally, AI is becoming more powerful and significantly cheaper over time (measured in months), making tasks previously unachievable due to high computational costs now widespread (Suleyman, 2023). In other words, AI is and will continue to bring an enormous amount of power to human civilization, to the extent that Sundar Pichai, CEO of Google, believes its significance surpasses even fire and electricity (Clifford, 2018).

Along with the immense potential of AI come new security vulnerabilities and more complex information security risks. As society is still transitioning to adapt to technological changes, the AI revolution continues, requiring us to rethink the trajectory of societal transformation because it could exacerbate the information security risks presented above. Specifically, how does society need to evolve to keep pace with the breakthrough changes brought about by the current wave of AI technology? How can we manage and leverage its power while ensuring information security as our living space and time in the virtual world continue to expand?

To contribute to the answers to these questions, the next section of this article will discuss the issues and risks affecting information security in the AI era, as well as the role, advantages, and opportunities of applying AI for the purpose of ensuring information security.

Following that, the way humans interact with AI, and the impact of personal AI use rights on information security in cyberspace will be examined, laying the foundation for discussions on the roles of governments, businesses, and citizens in ensuring information security. Some implications for improving information security are eventually provided, with an emerging country (i.e., Vietnam) being an exemplary context.

2. The Era of Artificial Intelligence and Its Impact on Cybersecurity

2.1. The Impact of AI on Attack and Defense Activities

Artificial Intelligence (AI) technology has demonstrated its superior potential in automating tasks, making predictions, and enhancing efficiency. As a result, AI has revolutionized the field of information security. Information security involves the management, monitoring, and protection activities carried out to minimize information risks. For the protection and defense of personal information, computer systems, and critical infrastructure, the main focus is to achieve the CIA triad while ensuring the efficient operation of the protected systems. The CIA triad includes (Maalem Lahcen et al., 2020):

- Confidentiality (C): Protecting data and systems from risks arising from data theft activities targeting databases, backups, application servers, and management systems.
- Integrity (I): Protecting data and systems from risks affecting the integrity of information and management systems, including hijacking control, altering financial data, stealing money, diverting stored information, and harming the organization's brand.
- Availability (A): Protecting data and systems from Denial of Service (DDoS) attacks, targeted Denial of Service attacks, and physical destruction risks.

The advent of AI has simultaneously increased the cyber attack capabilities of hackers and the defensive and security capabilities of network administrators significantly. Thanks to the ability to automate repetitive tasks and avoid human cognitive blind spots, machine learning algorithms can analyze vast amounts of information to identify security vulnerabilities that were previously undetectable (Rao, 2021). From a defensive perspective, the task of reviewing and searching for security vulnerabilities previously took a lot of time and effort due to the large number of recorded security flaws. Finding unpatched vulnerabilities often relied heavily on the experience of white-hat hackers, security technicians, and vulnerability scanning tools. This led to systems not being thoroughly reviewed and patched, making them quickly discovered and exploited by hackers. AI-based tools can now be used to automate the process of identifying these vulnerabilities in software systems, networks, and other digital assets before hackers find and exploit them.

Additionally, AI-powered tools make attacks increasingly diverse and sophisticated. Cybercriminals use a variety of AI-based tactics to infiltrate personal information systems and company networks, such as:

- Developing advanced malware and ransomware.
- Conducting stealth attacks.
- Using AI to guess complex passwords and break CAPTCHA.
- Creating deepfake content and impersonating individuals on social media platforms.
- Utilizing AI frameworks to attack vulnerable systems.
- Leveraging Machine Learning (ML) to enhance penetration testing.

AI-based tools can also be used to launch targeted hybrid attacks specifically designed for individuals or organizations (Handa et al., 2019). These enable cybercriminals to infiltrate and hide within a company's network for extended periods to carry out stealth attacks. During this time, they can establish secret access points to an organization's critical infrastructure. While preparing to launch a broader attack, these criminals could intercept communications, steal data, spread harmful software, create accounts with high-level access to infiltrate other systems, or deploy ransomware.

Similarly, phishing attacks have become more sophisticated with AI's help. Now, it is easy to receive a fake email, a phone call, or even a video call, impersonating banks, government agencies, or even relatives. AI-generated deepfake information can perfectly mimic the security protocols of regulatory bodies or replicate the voice and behavior of impersonated individuals.

Conversely, AI's ability to learn and predict current and future situations makes it a potent tool for updating, developing, and adapting to changes in cybercriminals' attack methods. For example, AI's capability to analyze and detect malware. Over the past few decades, malware has evolved rapidly, leading to advanced malicious software capable of altering its structure/code with each infection (such as polymorphic and metamorphic malware) (Sharma & Sahay, 2014). This allows them to breach traditional security barriers like firewalls and disable intrusion detection systems. To combat this, AI technologies are becoming increasingly popular because they not only help detect malware but also predict and update knowledge about new or unclear malware forms (Rieck et al., 2011). Besides analyzing and detecting malware, AI is also being developed to recognize and counter phishing attacks, spam, intrusions into traffic management systems, and attacks on electrical systems and industrial control systems (Handa et al., 2019; Martínez Torres et al., 2019).

2.2. Some Limitations of AI

Although AI is regarded as a leading solution for the increasing need for information security, it also has some limitations. Firstly, the cost required to develop a bespoke AI system for security needs must be mentioned. While not entirely accurate, we can refer to OpenAI's ChatGPT-3 model as an example. Analysts and technologists estimate that training a language model like ChatGPT-3 could consume over 4 million USD (Vanian & Leswing, 2023). Moreover, to undertake this training process, a company must have access to the necessary experts, machinery, data, and databases. This is almost beyond the reach of most individuals and small and medium-sized enterprises.

Of course, the cost of using AI models provided by technology companies will be much lower. For example, Microsoft offers the security system Copilot. This software is developed based on GPT-4, the largest current language model from OpenAI – in which Microsoft has invested billions of USD – and a specific security model that Microsoft has built by using the operational data it collects daily (Novet, 2023). Microsoft plans to charge a fee of 4 USD for each "security compute unit," and users can buy only what they need for their security requirements (Novet, 2024). However, this lower cost comes with another information security risk: the user's security environment information will be collected by technology companies. Microsoft itself has admitted that: "The [Copilot] system will know about the customer's security environment, but that data will not be used to train models" (Novet, 2024). While Microsoft commits not to use the collected data for "model training" purposes, they did not specify other purposes beyond "model training." If users and businesses do not care about this because their operations are not affected, the information collected from millions of users and hundreds of thousands of companies will be valuable for espionage and manipulation activities on a national and regional scale. It is frightening to think that we pay to enhance security yet allow the security service provider to know all the vulnerabilities in our systems.

Additionally, as AI is more widely applied to security work, more non-traditional security vulnerabilities will emerge. AI provides the ability to make automatic and continuous decisions over long periods, helping to detect malware or anomalies in the system. However, to do this, AI must be trained to differentiate signs of malware or abnormalities. Cybercriminals can exploit this training phase to adjust the output of the classification model, thereby manipulating the AI system to allow malware or malicious code into the system (Biggio, Fumera, et al., 2013; Handa et al., 2019). These types of attacks can be divided into two categories (Biggio, Corona, et al., 2013):

- Poisoning attacks: The attacker affects the training data, changes the training process, and damages the classification performance of AI.
- Evasion attacks: The attacker uses strategies to probe or perform offline analysis to find information that helps them manipulate the judgment of the classification system without having to impact the AI's training process.

While AI can provide powerful solutions for security purposes, it is not infallible. AI still needs to be controlled and governed by users, so security systems will always have potential vulnerabilities caused by human error. These human errors can be classified based on the consequences and intentions of the actor (Maalem Lahcen et al., 2020):

- Unintentional human error: Errors arising from a lack of knowledge or operational skills.

- Intentional human error: Errors caused by a user who is aware of the risky behavior but still acts on it or misuses the system. Such actions do not necessarily cause immediate loss to the organization but can still violate current laws or privacy rights.
- Malicious human error: The worst type of error because it is committed with the specific intent to harm the system.

Since the operators and controllers of data and systems are outside the AI system's scope of control, security vulnerabilities can still arise from deliberate sabotage behaviors within the internal team (or the operators themselves) (Maalem Lahcen et al., 2020). Sometimes human decisions and behaviors are irrational and unpredictable, influenced by anger, frustration, and job dissatisfaction, leading them to carry out intentional sabotage (malicious error), unsafe interventions (intentional error), or commit "naive" errors due to carelessness (unintentional error), etc. (Stanton et al., 2005). According to the 2023 Insider Threat Report, 74% of the surveyed cybersecurity experts feel that data and system security is vulnerable to internal threats. Furthermore, 74% of respondents also mentioned that insider attacks have become more frequent in the past 12 months (Insiders, 2023).

3. As Artificial Intelligence Becomes Stronger, the Human Factor Becomes More Important

Over the past decade, AI has developed rapidly and achieved breakthroughs, surpassing human capabilities in various fields and tasks (Henshall, 2023), including some aspects of information security. Although AI's computational abilities are increasing, the functions or products created by AI are still directed and decided by humans (through model training processes and direct commands to AI). In other words, the faster AI develops, the more it amplifies the capabilities and power of its users (or those it serves). This can lead to two issues.

Firstly, the power AI brings provides users with more choices. Tasks that were previously unattainable due to knowledge, ability, strength, and time limitations can now be supplemented by AI, requiring only learning to control AI effectively. However, why is this a problem? The issue is that they might choose to use AI for malicious purposes, such as conducting scam attacks, ransom attacks, creating malware, etc. This contributes to creating more information security risks in the future when someone with no security expertise can quickly become a black-hat hacker if they know how to control AI for cyber-attack purposes. The consequences could be worse if such hackers suddenly emerge from within an organization or company.

Moreover, the greater the power AI provides to users, the more significant their impact on everything around them. As previously discussed, despite AI's extremely powerful security capabilities, risks from human-caused vulnerabilities will always exist. If humans err in operating AI, the consequences of that error could be much more severe. For instance, human errors, whether accidental or intentional, that make the training data for AI's classification model inaccurate could exponentially increase the security risks due to AI's principle of continuous automatic operation over a long period, with human intervention being difficult (and costly if feasible). If AI continues to develop at a rate exceeding all predictions as it is now, its integration into every aspect of daily life for individuals, businesses, and nations, both in-depth and breadth, could soon become a reality (Henshall, 2023; Stacey & Milmo, 2023). Especially with the emergence of cyber-physical systems, like smart grids, autonomous vehicles, medical monitoring, industrial control systems, robotics, etc., the gap between the real and virtual worlds will continue to narrow. Then, impacts in the virtual world will have the potential to affect the real world directly. A single error in AI's automatic operation process caused by humans (especially in security issues) could lead to severe, unforeseeable consequences.

Both of the above issues stem from a change in the structure of power within society (Suleyman, 2023). Here, power can be understood as "the ability to create or prevent change" (Green, 1998). Therefore, to achieve information security in the AI era, we need a clearer understanding of the human and societal factors in the development and operation of AI, especially issues related to individual's freedom, power, and responsibility, the role of regulatory organizations and the state, and the responsibility of technology companies.

4. Social Structure Shift and the Concept of Freedom in the AI Era

To better understand the role of human factors and social structure in information security in the AI era, it is necessary to consider from the most fundamental components of social structure: individual thoughts, decisions, and behaviors. As the social structure is shifting from a phase without AI to a phase where AI is integrated into all aspects of life, the Mindsponge theory is used to clarify the issue thanks to its dynamic explanation capability centered around interaction with information.

The Mindsponge theory posits that each individual is a biological information storage and processing system (or an information collection-cum-processor) capable of making decisions and behaviors to interact with their environment (including natural, social, cultural, political, and technological environments) (Vuong, 2023). The operation of the information processing system includes the process of evaluating costs and benefits with the goal of optimizing perceived benefits and minimizing perceived costs (Vuong et al., 2022). These cost and benefit evaluations are influenced by the objectives and priorities of the system, as well as following the principle of energy conservation of organisms. The most basic purpose or priority of the system is to ensure the prolongation of the system's existence in one way or another, including survival, growth, and reproduction (Vuong, 2023). Through the Mindsponge information processing lens, we can envision that each individual's perception of power (the perception of the ability to create or prevent change/impact) is the product of information processing and interaction with their surrounding environment. Perceptions of power have limitations due to observations from objective reality and individual subjective evaluations related to knowledge, capability, strength, assets, social status, and time (Nguyen et al., 2023).

As AI begins to emerge and is applied in society, individuals will gradually observe the benefits that AI brings from an objective reality and choose to use them. Through the process of interaction and information exchange with AI, the initial perceptions (before knowing AI) gradually transform. These perceptions include those of the individual's own limits in knowledge, skills, strength, and time. With AI, individuals now have the ability to do things they previously could not or did not think of due to objective limitations in knowledge, skills, strength, and time. For example, someone who never knew how to draw or about computer programming can now easily create artistic images or computer programming codes by leveraging AI. Moreover, AI Deepfake now gives them the power to quickly and easily create realistic fake content, such as fake images and videos of other people's faces and voices.

When the objective power (or the ability to create or prevent change) (Green, 1998) of individuals is rapidly increased with the help of AI, it means the set of possible actions for that individual is also increasing. In other words, the overall freedom of the individual increases (Pansardi, 2012). Without accurate management mechanisms, this can significantly increase information security risks (as explained in Section 3).

In reality, an individual's overall freedom in society is limited by social systems. Although they have the capacity to perform a set of actions, due to the prevention or influence of other individuals or groups in society (through laws, rules, culture, or ethics), they do not perform some of the actions they are capable of (Kramer, 2008; Pansardi, 2012). From the Mindsponge theoretical perspective, the individual has the objective capability to perform actions but does not do so due to their subjective cost evaluations (created by others through laws, rules, culture, or ethics) (Nguyen et al., 2023; Vuong, 2023).

Currently, as the emergence of AI in life is still new and its future development remains uncertain, cultural norms and ethics around AI usage remain controversial and undefined. Meanwhile, the world's first law on artificial intelligence management was only approved by the European Union on August 13, 2024 (Liaropoulos, 2020). Therefore, we need a deeper understanding of the shift in social structure due to the change in power and the level of freedom that AI brings on a large scale to be able to deploy appropriate mechanisms to control power and regulate freedom.

Usually, these power control mechanisms are managed and deployed by the state. But why would individuals agree to lose some of their freedom, or in other words, allow the state to limit their own power?

This can be explained through the Social Contract Theory (Hobbes, 1894; Locke, 1967; Rousseau, 2016). This theory suggests that individuals collectively form a body with authority (e.g., the state) and relinquish a portion (even all under certain severe conditions concerning the survival of the social collective) of their freedom to this entity to manage and fulfill their responsibilities as described in the law. In return, the authority must provide those individuals within the collective with the benefits of political and social order, such as stability, personal safety, and property (Bierens et al., 2017; Boucher & Kelly, 2003; Liaropoulos, 2020). With the emergence of private companies in the 20th century, a third party was added to the social contract (Liaropoulos, 2020). They are seen as a legal entity in a country with the goal of maximizing profit, through which they create a development incentive for society (e.g., creating jobs, wealth, promoting innovation, etc.). However, private companies are not allowed to harm the social contract between citizens and the state; hence, the state has the right to apply specific laws and regulations to private companies while considering other factors, such as market competition between companies and the public. If a company becomes a monopoly or near-monopoly, government laws and regulations need to be strengthened by the state to control (Bierens et al., 2017; Liaropoulos, 2020).

However, the information revolution, and most recently, the emergence of AI, has made the world hyper-connected and changed the power structure in society by enhancing the power of those who can access and utilize AI. This leads to the question of whether national governments are capable of maintaining political order and social stability. If so, to what extent and scope, since the virtual space is almost without borders? Conversely, when individuals have gained the unimaginable power of AI, meaning their overall freedom has increased widely, are they still willing to trade that freedom for social stability as before? If so, how much freedom are they willing to trade to optimize the benefits they perceive? What happens if community behavior rules shared on information platforms become conflicted with intrinsic social contracts, eroding ethical systems and becoming super rules capable of causing widespread super-cultural conflicts?

Moreover, currently, governments do not have effective tools to limit the power of users multiplied by AI and other information technologies, as the main providers of these services are leading multinational technology corporations, such as Microsoft, Meta, Google, etc. More profoundly, these corporations hold most of the digital assets (data, software) and the infrastructure to operate digital technologies and AI (Nilekani, 2018). Most internet search data is stored by Google, while Meta (previously Facebook) dominates social networking with over 2 billion users. With the vast number of users and the huge amount of data obtained from them, although these conglomerates do not own much physical property, have no police, courts, or similar state institutions, they still have the ability to control information sources, influence opinions, and even manipulate the psychology and behavior of a large number of users (Shadmy, 2019).

In an era of exploding information technology and AI, the change in the power structure of society's components is happening. The transformation, even an upgrade, in the social contract is necessary for society to adapt and even evolve, but it must also ensure political and social stability, within which information security is an essential part. Social contracts that only involve individual governments are unlikely to be sustainable. Therefore, the social contract needs coordination and connection among parties through cooperation between governments, supranational organizations, public-private partnerships, citizens, nongovernmental organizations, and private companies (especially technology conglomerates) (Liaropoulos, 2020).

5. Awareness, Investment in Cybersecurity, and Some Recommendations

Cybersecurity and information security play a crucial role in modern socio-economic activities and national protection (Nash-Hoff, 2012). In the context of globalization and economic integration, the relationship between the economy, especially e-commerce, and national security has become increasingly intertwined (Okhrimenko et al., 2023). As technology advances and the space and time people spend in the virtual world increase, with cyber-physical systems being deployed and operated more broadly in the global economy and social activities, protecting information becomes an essential need to ensure not only information security but also sustainable development and national security.

The rapid development of AI has shown superior potential in the field of information security, but it also brings significant concerns, as hackers could also use AI for cyber-attacks or fraud. This invisibly creates a race between the defense and the attacking sides: whoever can develop better, faster, and more effectively utilized AI will have more advantages. Therefore, focusing resources on developing AI for use in cybersecurity needs to be a priority investment to ensure the countries', businesses', and individuals' assets (information) are not lost or exploited illicitly for malicious purposes or espionage. However, investment efficiency issues also need to be carefully considered to avoid ineffective investment and waste (Vuong, 2018).

Investing in new AI models will be very costly and beyond the affordability of most businesses, especially in emerging countries like Vietnam. Moreover, AI is a machine learning system, so it needs continuous training and updating with new features and algorithms to ensure the system can respond to increasingly sophisticated and customized attack methods of cyber criminals. Using AI models developed by large technology corporations, like Microsoft, will significantly reduce security-related costs. However, this approach will expose all security weaknesses to the AI service providers. If this happens on a large scale, it could lead to espionage and manipulation risks at the national level. Therefore, the government needs specific support policies and programs to collaborate with domestic cybersecurity businesses to develop their own AI security systems, alongside using external service providers for types of data and information systems that do not significantly affect national security. Faced with national security challenges, this collaboration fundamentally has to eliminate purely commercial conflicts of interest while still ensuring legitimate interests and intellectual property rights.

Currently, information safety and security in Vietnam are witnessing significant advancements. In 2023, Vietnam aims to become a "cybersecurity powerhouse" by 2025, focusing on the development and export of cybersecurity products and services. The country is also concentrating on building a high-quality workforce in this field (Anh, 2024). Vietnam has cybersecurity companies and organizations capable of providing professional security and information safety services, such as Viettel, Vietnam Cybersecurity Technology JSC, HPT

Information Technology Services JSC, CMC Technology Group JSC... Moreover, Vietnam is a participant in the World Online Authentication Alliance (FIDO), accessing advanced non-password authentication technologies and trends, and hosting the International Cybersecurity Day Vietnam 2023 Conference and Exhibition (Tạp chí An toàn thông tin, 2023). These are very useful prerequisites for Vietnam to deploy and develop a dedicated AI model for security.

From a business perspective, awareness of security and information safety in enterprises is not yet profound. Most Vietnamese companies still use an “ad-hoc” IT team for both system development and security tasks instead of hiring professional entities. Efforts to prevent malicious information are weak and flawed. The maturity level of cybersecurity in businesses is not commensurate with the threat to information safety. The recent cyber attack on VNDirect, one of Vietnam’s leading securities companies, which resulted in their system being taken down, reveals a concerning truth: awareness about information security and safety among many businesses remains significantly limited. This situation has led to a scenario where the economic damage suffered by the company post-incident is vastly greater than what the cost would have been for investing in professional information security measures from the outset (vnexpress.net, 2024). A 2021 McKinsey cybersecurity maturity survey of over 100 companies across various sectors showed a correlation between cybersecurity maturity levels and profit margins. This indicates that effective cybersecurity strategies can contribute to the overall financial health of a company (Eiden et al., 2021). Therefore, Vietnamese companies need to be more serious about investing in cybersecurity measures, especially in the AI era, where cybercriminals can quickly develop both in quantity and quality.

As AI becomes stronger and more versatile, the human factor becomes extremely important because it will help determine the effectiveness of AI applications and resilience against information security risks. Therefore, in addition to investing in developing AI models for security purposes, activities to raise awareness about the importance of information and the risks of information exploitation, as well as training and educating the public, businesses, and government agencies on how to protect information and information systems also need to be emphasized. In this way, individuals, businesses, and government agencies participating in cyberspace activities will have the awareness and ability to protect themselves against security risks, thereby contributing to the sustainability of the national information space. Indeed, information security issues are increasingly prevalent in Vietnam. Examples include recent scams on applications like Zalo and Telegram, or the emergence of deepfake technology in fraud cases (Sơn, 2023).

No matter how perfect an AI-integrated security system is, it will always have the potential to overlook vulnerabilities caused by human error. One notable issue is the lack of full compliance with information safety regulations by some government agencies. This is evident when simply searching for keywords like “gambling” or “football” on state agency domains, which can reveal hacker intrusions and the appearance of unwanted content. These security incidents not only lead to the dissemination of inappropriate information but also pose a significant risk if hackers exploit them to disseminate false information or engage in fraudulent activities, causing serious consequences. This issue requires timely remedial measures to enhance the safety and security of government agencies' information systems.

Vietnam also needs to recognize the importance of developing human resources in the field of information technology and cybersecurity (Vuong et al., 2019). There are already some universities actively incorporating Information Security subjects into their curricula. However, both the quantity and quality of these courses have not yet truly met the demand and are still at a preliminary stage. To keep pace with the rapid development of technology, the content of these courses, along with the faculty, needs to be continuously updated to meet new technological advancements and adapt to current trends. This not only provides necessary knowledge and skills to students but also contributes to enhancing the overall capacity of the information technology and cybersecurity sectors in Vietnam. Additionally, the government and universities should guide and promote social science, psychology, and behavioral research related to cybersecurity issues, as humans remain the most crucial and also the most vulnerable link in achieving comprehensive information security goals. Currently, the number of studies on this issue remains limited (Maalem Lahcen et al., 2020; Payne & Hadzhidimova, 2018).

In summary, strengthening compliance with information safety regulations, countering new forms of scams, and improving awareness of information security within the business community are areas that need attention. At the same time, developing high-quality human resources in this field, especially through updated and specialized university teaching programs, will be key for countries, especially emerging ones like Vietnam, to strengthen their socio-economic security and continue to advance in the AI era.

References

Anh, N. (2024). *Hướng tới mục tiêu "cường quốc an toàn thông tin mạng"*. VnEconomy. Retrieved March 19 from https://vneconomy.vn/huong-toi-muc-tieu-cuong-quoc-an-toan-thong-tin-mang.htm

Bierens, R., Klievink, B., & van Den Berg, J. (2017). A social cyber contract theory model for understanding national cyber strategies. Electronic Government: 16th IFIP WG 8.5 International Conference, St. Petersburg, Russia.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., . . . Roli, F. (2013). Evasion attacks against machine learning at test time. Machine Learning and Knowledge Discovery in Databases: European Conference, Prague, Czech Republic.

Biggio, B., Fumera, G., & Roli, F. (2013). Security evaluation of pattern classifiers under attack. *IEEE Transactions on Knowledge and Data Engineering*, *26*(4), 984-996. https://doi.org/10.1109/TKDE.2013.57

Boucher, D., & Kelly, P. (2003). *The social contract from Hobbes to Rawls*. Routledge.

Chính, P. M., & Hoàng, V. Q. (2009). *Kinh tế Việt Nam: Thăng trầm và đột phá*. Nxb Chính trị quốc gia-Sự thật.

Clifford, C. (2018). *Google CEO: A.I. is more important than fire or electricity*. CNBC. Retrieved March 18 from https://www.cnbc.com/2018/02/01/google-ceo-sundar-pichai-ai-is-more-important-than-fire-electricity.html

Eiden, K., Kaplan, J., Kazimierski, B., Lewis, C., & Telford, K. (2021). *Organizational cyber maturity: A survey of industries*. https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/organizational-cyber-maturity-a-survey-of-industries

Green, L. (1998). Power. In *Routledge Encyclopedia of Philosophy*: Taylor and Francis.

Greenberg, A. (2015). *Hackers remotely kill a Jeep on the highway—With me in it*. WIRED. Retrieved March 17 from https://www.wired.com/2015/07/hackers-remotely-kill-jeep-highway/

Handa, A., Sharma, A., & Shukla, S. K. (2019). Machine learning in cybersecurity: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(4), e1306. https://doi.org/10.1002/widm.1306

Henshall, W. (2023). *4 charts that show why AI progress is unlikely to slow down*. Time. Retrieved March 18 from https://time.com/6300942/ai-progress-charts/

Ho, M.-T., & Vuong, Q.-H. (2023). Disengage to survive the AI-powered sensory overload world. *AI and Society*. https://doi.org/10.1007/s00146-023-01714-0

Hobbes, T. (1894). *Leviathan: Or, the matter, form, and power of a commonwealth ecclesiastical and civil* (Vol. 21). G. Routledge and sons.

Hu, K. (2023). *ChatGPT sets record for fastest-growing user base - analyst note*. Reuters. Retrieved May 11 from https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Insiders, C. (2023). *2023 insider threat report*. https://istari-global.com/insights/spotlight/2023-insider-threat-report/

Keck, M., Gillani, S., Dermish, A., Grossman, J., & Rühmann, F. (2022). *The role of cybersecurity and data security in the digital economy*. https://policyaccelerator.uncdf.org/all/brief-cybersecurity-digital-economy

Kramer, M. H. (2008). *The quality of freedom*. Oxford University Press.

Lappin, S. (2023). Assessing the strengths and weaknesses of Large Language Models. *Journal of Logic, Language and Information*, *33*, 9-20. https://doi.org/10.1007/s10849-023-09409-x

Liaropoulos, A. (2020). A social contract for cyberspace. *Journal of Information Warfare*, *19*(2), 1-11. https://www.jstor.org/stable/27033617

Locke, J. (1967). *Two treatises of government*. Cambridge university press.

Lu, Y., & Da Xu, L. (2018). Internet of Things (IoT) cybersecurity research: A review of current research topics. *IEEE Internet of Things Journal*, *6*(2), 2103-2115. https://doi.org/10.1109/JIOT.2018.2869847

Maalem Lahcen, R. A., Caulkins, B., Mohapatra, R., & Kumar, M. (2020). Review and insight on the behavioral aspects of cybersecurity. *Cybersecurity*, *3*, 1-18. https://doi.org/10.1186/s42400-020-00050-w

Mantello, P., Ho, M.-T., Nguyen, M.-H., & Vuong, Q.-H. (2023). Machines that feel: behavioral determinants of attitude towards affect recognition technology—upgrading technology acceptance theory with the mindsponge model. *Humanities and Social Sciences Communications*, *10*, 430. https://doi.org/10.1057/s41599-023-01837-1

Marr, B. (2021). *How much data do we create every day? The mind-blowing stats everyone should read*. Bernard Marr & Co. Retrieved March 18 from https://bernardmarr.com/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/

Martínez Torres, J., Iglesias Comesaña, C., & García-Nieto, P. J. (2019). Machine learning techniques applied to cybersecurity. *International Journal of Machine Learning and Cybernetics*, *10*(10), 2823-2836. https://doi.org/10.1007/s13042-018-00906-1

Morgan, S. (2022). *Cybercrime to cost the world 8 trillion annually in 2023*. Cybercrime Magazine. Retrieved March 18 from https://cybersecurityventures.com/cybercrime-to-cost-the-world-8-trillion-annually-in-2023/

Musa, S. (2018). Smart cities-a road map for development. *IEEE Potentials*, *37*(2), 19-23. https://doi.org/10.1109/MPOT.2016.2566099

Nash-Hoff, M. (2012). *What does the economy have to do with national security?* IndustryWeek. Retrieved March 19 from https://www.industryweek.com/finance/software-systems/article/21954333/what-does-the-economy-have-to-do-with-national-security

Nguyen, M.-H., Le, T.-T., & Vuong, Q.-H. (2023). Ecomindsponge: A novel perspective on human psychology and behavior in the ecosystem. *Urban Science*, *7*(1), 31. https://doi.org/10.3390/urbansci7010031

Nilekani, N. (2018). Data to the people: India's inclusive internet. *Foreign Affairs*, *97*(5), 19-27.

Novet, J. (2023). *Microsoft introduces an A.I. chatbot for cybersecurity experts*. CNBC. Retrieved March 18 from https://www.cnbc.com/2023/03/28/microsoft-launches-security-copilot-in-private-preview.html

Novet, J. (2024). *Microsoft says new AI security chatbot pricing model lets customers 'buy what they need'*. CNBC. Retrieved March 18 from https://www.cnbc.com/2024/03/13/microsoft-uses-compute-units-to-charge-customers-for-security-copilot.html

Okhrimenko, I., Stepenko, V., Chernova, O., & Zatsarinnaya, E. (2023). The impact of information sphere in the economic security of the country: case of Russian realities. *Journal of Innovation and Entrepreneurship*, *12*(1), 67. https://doi.org/10.1186/s13731-023-00326-8

Paiho, S., Tuominen, P., Rökman, J., Ylikerälä, M., Pajula, J., & Siikavirta, H. (2022). Opportunities of collected city data for smart cities. *IET Smart Cities*, *4*(4), 275–291. https://doi.org/10.1049/smc2.12044

Pansardi, P. (2012). Power and freedom: opposite or equivalent concepts? *Theoria*, *59*(132), 26-44. https://www.jstor.org/stable/41802526

Payne, B. K., & Hadzhidimova, L. (2018). Cyber security and criminal justice programs in the United States: Exploring the intersections. *International Journal of Criminal Justice Sciences*, *13*(2). https://doi.org/10.5281/zenodo.2657646

Rao, Vikram Singh. (2021). *Best AI-based cyber security tools for improved safety*. Echnotification. Retrieved March 19 from https://www.technotification.com/2021/06/best-ai-based-cyber-security-tools.html

Rieck, K., Trinius, P., Willems, C., & Holz, T. (2011). Automatic analysis of malware behavior using machine learning. *Journal of Computer Security*, *19*(4), 639-668. https://doi.org/10.5555/2011216.2011217

RiskXchange. (2023). *Cybersecurity statistics you should know in 2023*. RiskXchange. Retrieved March 19 from https://riskxchange.co/1006415/cybersecurity-statistics-2023/

Rousseau, J.-J. (2016). The social contract. In R. Blaug & J. Schwarzmantel (Eds.), *Democracy: A Reader* (pp. 43-51). Columbia University Press.

Shadmy, T. (2019). The new social contract: Facebook's community and our rights. *Boston University International Law Journal*, *37*, 307.

Sharma, A., & Sahay, S. K. (2014). Evolution and detection of polymorphic and metamorphic malwares: A survey. *International Journal of Computer Applications*, *90*(2), 7–11. https://doi.org/10.48550/arXiv.1406.7061

Sơn, M. (2023). *An ninh mạng: Những xu hướng đáng chú ý trong 6 tháng cuối năm 2023*. VietnamPlus. Retrieved March 19 from https://www.vietnamplus.vn/an-ninh-mang-nhung-xu-huong-dang-chu-y-trong-6-thang-cuoi-nam-2023-post869804.vnp

Stacey, K., & Milmo, D. (2023). *AI developing too fast for regulators to keep up, says Oliver Dowden*. The Guardian. Retrieved March 18 from https://www.theguardian.com/technology/2023/sep/22/ai-developing-too-fast-for-regulators-to-keep-up-oliver-dowden

Stanton, J. M., Stam, K. R., Mastrangelo, P., & Jolton, J. (2005). Analysis of end user security behaviors. *Computers and Security*, *24*(2), 124-133. https://doi.org/10.1016/j.cose.2004.07.001

Suleyman, M. (2023). *How the AI revolution will reshape the world*. Time. Retrieved March 18 from https://time.com/6310115/ai-revolution-reshape-the-world/

Tạp chí An toàn thông tin. (2023). *An toàn thông tin 10 dấu ấn nổi bật trong lĩnh vực bảo mật và an ninh, an toàn thông tin tại Việt Nam năm 2023*. Trung tâm Công nghệ Thông tin và Truyền thông Nghệ An. Retrieved March 19 from https://naict.tttt.nghean.gov.vn/attt/an-toan-thong-tin-10-dau-an-noi-bat-trong-linh-vuc-bao-mat-va-an-ninh-an-toan-thong-tin-tai-viet-nam-nam-2023-597.html

VnExpress.net. (2024). *VNDirect associate companies under cyberattack*. VnExpress. Retrieved March 25 from https://e.vnexpress.net/news/business/companies/vndirect-associate-companies-under-cyberattack-4726352.html

Vailshery, L. S. (2023). *Number of IoT connected devices worldwide 2019-2023, with forecasts to 2030*. Statista. Retrieved March 18 from https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/

Vanian, J., & Leswing, K. (2023). *ChatGPT and generative AI are booming, but the costs can be extraordinary*. CNBC. Retrieved March 18 from https://www.cnbc.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html

Vuong, Q.-H. (2018). The (ir) rational consideration of the cost of science in transition economies. *Nature Human Behaviour*, *2*(1), 5.

Vuong, Q.-H. (2023). *Mindsponge Theory*. Walter de Gruyter GmbH. https://www.amazon.com/dp/B0C3WHZ2B3/

Vuong, Q.-H. (2022). *The Kingfisher Story Collection*. https://www.amazon.com/dp/B0BFV9L58W

Vuong, Q.-H., *et al*. (2019). Artificial intelligence vs. natural stupidity: Evaluating AI readiness for the vietnamese medical information system. *Journal of Clinical Medicine*, *8*(2), 168. https://doi.org/10.3390/jcm8020168

Vuong, Q.-H., *et al*. (2023a). AI's humanoid appearance can affect human perceptions of Its emotional capability: Evidence from self-reported data in the US. *International*

Journal of Human–Computer Interaction, 1-12. https://doi.org/10.1080/10447318.2023.2227828

Vuong, Q.-H., *et al*. (2023b). How AI's self-prolongation influences people's perceptions of its autonomous mind: The case of US residents. *Behavioral Sciences*, *13*(6), 470. https://doi.org/10.3390/bs13060470

Vuong, Q.-H., *et al*. (2023). Are we at the start of the artificial intelligence era in academic publishing? *Science Editing*, *10*(2), 158-164. https://doi.org/10.6087/kcse.310

Vuong, Q.-H., Nguyen, M.-H., & La, V.-P. (2022). *The mindsponge and BMF analytics for innovative thinking in social sciences and humanities*. Walter de Gruyter GmbH. https://www.amazon.com/dp/B0C4ZK3M74/

Wise, J. (2023). *How many Google searches per minute in 2024?* EarthWeb. Retrieved March 18 from https://earthweb.com/how-many-google-searches-per-minute/

World Economic Forum. (2023). *The global risks report 2023*. https://www3.weforum.org/docs/WEF_Global_Risks_Report_2023.pdf