Security and Safety, Vol. 2, 2023012 (2023)

Security and Safety in the “Metaverse”

Views OPEN ACCESS

Other Fields

On challenges of AI to cognitive security and safety

Ruiyang Huang[1], Xiaoqing Zheng[2], Yuting Shang[2], and Xiangyang Xue[2,*]

[1] *National Digital Switching System Engineering & Technological R&D Center, Zhengzhou 450002, China*
[2] *Fudan University, Shanghai 200433, China*

Received: 6 April 2023 / Revised: 30 April 2023 / Accepted: 30 April 2023 / Published online: 30 June 2023

Abstract Recent advances in deep learning have led to disruptive breakthroughs in artificial intelligence (AI), fueling the jump in ChatGPT-like large language models (LLMs). As with any emerging technology, it is a two-sided coin, bringing not only vast social impacts but also significant security concerns, especially in the socio-cognitive domain. Against this background, this work starts with an inherent mechanism analysis of cognitive domain games, from which it proceeds to explore the security concerns facing the cognitive domain as well as to analyze the formation mechanisms of a cognitive immune system. Finally, inspired by behavioral mimicry in biology, this work will elaborate on new approaches to cognitive security from three aspects: Mimicry Computing, Mimicry Defense, and Mimicry Intelligence.

Keywords Cognitive security, Artificial intelligence, Endogenous safety and security, Mimicry intelligence

Citation Huang RY, Zheng ZQ, Shang YT, et al. On challenges of AI to cognitive security and safety. Security and Safety 2023; **2**: 2023012. https://doi.org/10.1051/sands/2023012

1 Introduction

The recent rapid development of information network (IT) and artificial intelligence (AI) technologies has greatly accelerated the trend of Human Cyber Physical Ternary Fusion, making the physical domain, cyber domain, and cognitive domain more overlapping and intertwining. Meanwhile, remarkable breakthroughs in the ChatGPT-like artificial intelligence generated content (AIGC) technology [1–4] have radically reduced the production threshold of fake, biased, incorrect, and other harmful information. The resulting “Cognitive Fog” will affect social cognition, and then spark off social problems. Even worse, due to the aforementioned Human Cyber Physical Ternary Fusion, security challenges of AI in the cognitive (human) domain, if not appropriately handled, may entail spillovers to other domains to bring about systemic risks [5, 6]. It follows that it is extremely meaningful and urgent work to research the challenges of AI in the cognitive domain [7, 8].

2 Security concerns in the cognitive domain

In the information age, human communication is undergoing a complex and significant transformation. As offline social activity gradually gives way to online interaction, large social media platforms have become the main front of cognitive games [9] and the main channel of influencing people’s cognition. Sharing information, emotions, ideas, and thoughts among users will inevitably affect one’s original cognition. Therefore, external cognitive intervention and manipulation become easier to realize. For this reason, security problems in the cognitive domain will become more and more serious [10, 11].

2.1 Objective reasons for security concerns in the cognitive domain

Human society is entering a "third living space", namely the cognitive domain [12–14]. This new living space represents several features. First, human activities begin to break through the original time and space scales and boundaries to obtain external information. This entails the ability to discern truth from falsehood so that people can benefit from the abundance of information without falling prey to myths and misrepresentations. The common sense of "a clear head comes from an open mind" may not be the gospel anymore. Second, a large number of creative activities are moving online. The space where humans learn and reason is no longer limited to the brain but extends to virtual reality which in turn impacts the physical world. Third, physical reality and virtual reality represent bidirectional mapping. If cyberspace is a digital projection of human activities in the real world, then the cognitive domain is the materialization of human spiritual activities.

2.2 Internal logic of security concerns in the cognitive domain

The space where cognitive processes occur is moving from the "biological brain" to the "biological brain plus digital brain" [15]. Thus, the emerging cognitive domain can be accredited to the two-way interaction of the "biological brain" and the "digital brain". Moreover, the impacts of the "digital brain" on its biological counterpart are scaling up [16, 17]. This process can be divided into four stages. In the first stage, the perception process of human beings migrates to the "digital brain", and the task of observing the world is handed over to machines (*i.e.*, AI systems). In the second stage, the understanding process migrates to the "digital brain", and the task of analyzing the world is handed over to machines. In the third stage, the learning process migrates to the "digital brain", and the task of gaining insight into the world is handed over to machines. In the fourth stage, the decision-making and imagination processes undergo a digital transformation, and the task of changing the world is handed over to machines. The "digital brain" itself does not produce "new" data or "new knowledge" until the fourth stage. In this stage, the "digital brain" can not only give a presentation of human psychological, emotional, and cognitive activities as well as social public opinion but also accurately grasp the cognitive activities of human beings and execute fine interventions on human cognition [18].

2.3 Examples of security concerns in the cognitive domain

Future research on cognitive security risks [19] should fully consider the interconnection, interaction, and even partial substitution between the "biological brain" and the "digital brain". There are mainly three categories of security issues.

2.3.1 Security issues associated with cognitive processes

There are inherent defects in human thinking, namely cognitive biases. Cognitive biases stem from a kind of perceptual shortcuts that humans naturally take at all times. Specifically, in order to improve cognitive efficiency and make rapid responses, human brains will try to simplify complex problems by ignoring some information to reduce the burden of the cognitive process, or through "excessive consumption" of some information to spare itself the trouble of collecting more. These perceptual shortcuts will affect processes of search, understanding, selection, memory, *etc.* Although they can help human beings employ their limited cognitive ability to process unlimited information, they may also bring about cognitive biases, cognitive blind spots, and cognitive errors.

2.3.2 Security issues associated with cross-domain convergence

The physical domain and the cyber domain are connected through cyber-physical systems (CPS), the cyber domain and the cognitive domain are interweaved through cyber-human systems (CHS), and the three domains are interconnected through cyber-physical-social systems (CPSS). This cross-domain convergence makes it possible to influence the public's social cognition. It is indisputable that the current distributed, multi-dimensional, and interactive communication mode in our hyper-connected world advanced by the Internet not only provides everyone with a "microphone" to be heard and the convenience to be a "media producer", but also creates a sense of synchronicity and presence for the public. However, due to a lack of truth and difficulty in distinguishing between the true and the false, it is hard to guarantee cognitive autonomy and information accuracy, thus making the public easy targets of deception and exploitation.

2.3.3 Security issues associated with AI technologies

While AI offers many benefits, it also presents several security issues in the cognitive domain. Some of the major security issues caused by AI technologies in the cognitive domain are:

- Deepfakes: AI-generated images, videos, or audio recordings can create highly convincing fake content that is difficult to distinguish from real content. This can lead to misinformation, manipulation, and even blackmail.
- Misinformation and disinformation: AI algorithms can rapidly spread false information, amplifying the impact of misinformation and disinformation campaigns. This can undermine trust in institutions, disrupt political processes, and provoke social unrest.
- Bias and discrimination: AI systems can inadvertently learn and perpetuate biases present in the training data. This may lead to unfair treatment, perpetuation of stereotypes, and exacerbation of social inequalities.
- Erosion of privacy: AI technologies like facial recognition, natural language processing, and data mining can be used to identify, track, and profile individuals, potentially violating their right to privacy and enabling surveillance.
- Manipulation and influence: AI-driven algorithms can personalize and target content, making it easier for bad actors to manipulate people's thoughts, beliefs, and behaviors, potentially leading to radicalization or other harmful outcomes.
- Automated cyberattacks: AI can be used to develop more sophisticated and effective cyberattacks, such as generating phishing emails that are harder to detect or automating the discovery and exploitation of software vulnerabilities.
- Adversarial attacks: AI systems can be vulnerable to adversarial attacks, where an attacker carefully modifies input data to cause an AI system to make incorrect predictions or classifications, potentially compromising the system's integrity and reliability.
- Dependence on AI: As individuals and organizations become more reliant on AI systems for decision-making, there is a risk of over-reliance, leading to a decline in critical thinking skills and a reduced ability to adapt when AI systems fail or are unavailable.
- Diminished human agency: As AI takes over more cognitive tasks, there is a risk of reduced human agency, potentially leading to a decline in individual autonomy and decision-making capabilities.

Recently, the AI chatbot "ChatGPT" developed by OpenAI has dazzled the world. It can generate a human-like text that is difficult to differentiate the true from the false. Its outstanding performance in generating content is bound to affect the social cognition of the global public through the Internet, thus incurring incalculable consequences to cognitive security. Addressing these security issues will require a combination of technical solutions, policy interventions, and international cooperation to ensure AI technologies are developed and deployed responsibly and ethically.

3 Inherent mechanism analysis of cognitive game

Human beings are a social species that relies on cooperation to survive and thrive. Cooperation entails interaction, at the core of which is to understand each other's intentions through information sharing and accordingly adjust one's own behaviors. In the long-term evolution, the human brain has demonstrated the ability to form 4 kinds of cognitive perceptions: "the you in your own eyes", "the you in the eyes of others", "the others in your eyes" and "the others in their own eyes" [19]. Inevitable discrepancies between the cognitive gains and the objective reality on the one hand, and motives of cognitive synergy for group collaboration and division of labor on the other hand, all lead to cognitive biases such as Bandwagon Effect, Groupthink, Stereotype, Belief Bias, False Memory, *etc.* [20–23]. These deviations provide the

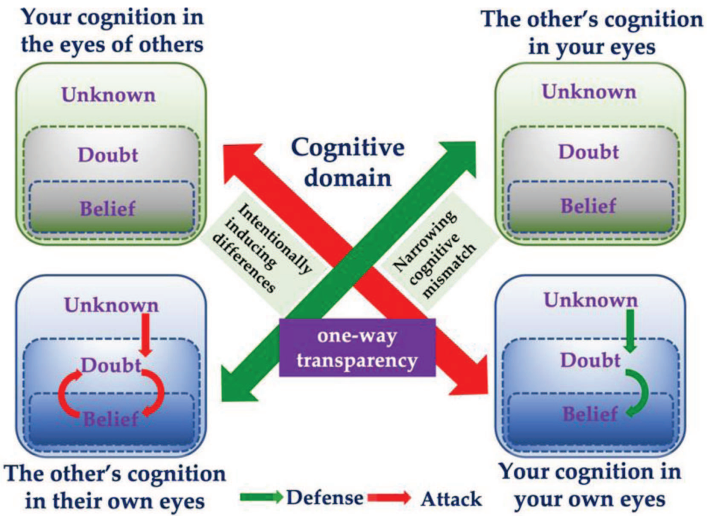
Your cognition in the eyes of others
The other's cognition in your eyes
Unknown
Doubt
Belief
Unknown
Doubt
Belief
Cognitive domain
Intentionally inducing differences
Narrowing cognitive mismatch
one-way transparency
Unknown
Doubt
Belief
Unknown
Doubt
Belief
The other's cognition in their own eyes
Defense
Attack
Your cognition in your own eyes

Figure 1. Conceptual map of the cognitive game.

possibility for the cognitive game or confrontation. A lot of cognitive psychology research has shown that there are indeed some brain sections with a prominent role in reconstructing images and intentions of self and others. Brain damage to these sections will not affect other functions or behaviors, except that it will result in the failure in understanding others' emotions and intentions and of carrying out collaboration.

3.1 How cognitive game works

The cognitive domain of a cognitive body is composed of such sub-spaces as the belief domain, doubt domain, and unknown domain. Although they do not overlap each other, their respective cognitive propositions can be transferred across them. The propositions that people believe in are located in the belief domain, the known but uncertain propositions are located in the doubt field, and the unknowns are located in the unknown domain. For example, the color of an object that has not been heard of or seen is unknown to people. When we are told that it is red, we may still have doubts. We become convinced that it's red when we have seen the object with our own eyes. But then, we are told that it might be illuminated by a red light when it is being observed. Once again, we begin to doubt whether it is red. People's cognition of an object's color may experience "twists and turns" under outside intervention. Similarly, in the face of enormous uncertain or inauthentic information on the Internet and social media platforms, human beings are prone to cognitive biases.

A cognitive Game or Cognitive Confrontation is focused on the cognitive domain of individuals, groups, and the wide public with a view to attacking, corroding and disrupting the targets' cognition mainly by means of strategic inducement, dissuasion and denial, information fog, *etc.* At least two parties (offense & defense) participate in this game (as shown in Figure 1). On the defense side, the task is to maximize the transparency of the opposite party's intentions by minimizing the mismatch between "the others in their own eyes" and "the others in your eyes" (analysis). On the offense side, the task is to minimize the transparency of one's own side by deliberately introducing deviations of "the you in the eyes of others" from "the you in your own eyes" (confrontation). It is in this way that the ultimate goal of "one-way transparency" in cognitive confrontation is achieved. Further, we can consciously publish carefully crafted information to change the distribution of cognitive propositions across the opposite party's unknown, doubt and belief domains, and carefully steer the distribution of those propositions across the three domains on one's own side, thereby occupying the high ground of cognitive confrontation.

3.2 Core competence of cognitive confrontation

The underlying logic of cognitive confrontation competence lies in the acquisition, processing, analysis, transmission, and utilization of information. In terms of system development and strategic competition, the focus shall be on the integration and confrontation of the "hide" and "seek" capabilities. The party that can get ahead to see through the cognitive space to identify the weakness of the other party (seek), and meanwhile deny its counterpart the same capability by creating more information fog, can effectively demoralize the other party and reduce the risk and cost of confrontation to the minimum, thus achieving the ultimate goal of subduing the enemy without fighting. Recent breakthroughs in AI technologies and deep learning in particular have made it possible for carrying out comprehensive game analysis in the cognitive domain based on cognitive modeling. The cognitive game in the post-ChatGPT era will largely be a machine-machine confrontation and a man-machine one. You can well imagine a fierce conflict unfolding itself in the use, channeling, intervention and control of data, information and knowledge in the cognitive domain hidden behind the Internet.

4 Endogenous safety and security problems in AI systems

With the wide application of AI technologies represented by deep learning, the ensuing security issues have attracted more and more attention from academia, the industry, and even society at large. Inherent security issues of AI systems, such as vulnerabilities, bugs, and backdoors, make them easy targets of exploitation to generate fake, biased, and incorrect information aimed at influencing public cognition. This will compound the security problems facing the cognitive domain. According to the Cyberspace Endogenous Safety and Security research perspective, the endogenous safety and security problems of AI systems can be subdivided into common problems, individual problems, and generalized functional safety (safety & security, S&S for short) problems.

4.1 Common endogenous security problems

Entities of AI systems are built on cyber-physical systems, so its "pedestal" of algorithm models will inevitably encounter common endogenous security problems. According to relevant research at home and abroad, "security bugs or defects" is a prevailing problem in software and hardware environments on which the mainstream deep learning framework relies. When these defects or bugs are exploited by attackers, AI systems are prone to disruption, manipulation, and data theft. With regard to software, China's 360 company conducted a security evaluation in 2021 on the prevalent open-source AI frameworks at home and abroad. It has found more than 150 vulnerabilities in seven machine learning frameworks (including TensorFlow, PyTorch, *etc.*) as well as another 200+ supply chain vulnerabilities. What about hardware? The University of California, Riverside has answered. In 2022, the university made public several security vulnerabilities in the NVIDIA GPU that will endanger user privacy.

4.2 Individual endogenous security problems

Individual endogenous security problems of AI systems refer to the security risks rooted in the vulnerabilities of AI algorithms and models. Existing studies have shown that a variety of attack methods can be employed by exploiting vulnerabilities in AI models, such as adversarial examples, data backdoors, data poisoning, *etc.* [24–29]. Adversarial examples are inputs of so delicately designed "subtle perturbations" by attackers to AI models that fail afterward to provide correct inference, thus leading AI systems to make a serious misjudgment. They are based on analysis and tests of the "vulnerable points" of AI models. Backdoor attacks involve creating backdoors in AI models by attackers who will activate them under certain conditions to carry out a malicious exploit. This kind of attack often goes undetected and thus it is difficult to find, locate and track them. Data poisoning involves tampering with machine learning training datasets and misleading AI algorithms to produce undifferentiated mistakes such that malicious examples will be labeled into desired ones or the performance of AI models will be greatly degraded, thereby damaging the availability of AI systems [30–35].

4.3 Generalized functional safety problems

While the above analysis of common and individual endogenous safety and security problems mainly focuses on clearing up the causes of security threats to AI applications, the subsequent analysis of generalized functional safety problems (security & safety, S&S for short) will focus on the consequences of security threats. As attention-grabbing deep learning technology is widely used in natural language processing, speech recognition, and computer vision, it is also gradually applied in manufacturing, transportation, medical treatment, people's livelihood, *etc.* Every trade takes on a new look of "intelligence + everything". With the further integration of AI with human life and production, generalized functional safety problems come to the foreground. Based on the individual problems or common problems of AI applications, attackers can ravage and maliciously tamper with the functions of the target system to incur dreadful effects.

In the future, adverse impacts caused by the interweaving and overlapping of cybersecurity, functional safety, and information security problems will only get more serious. Thus AI systems must be able to deal with random disturbances and man-made or natural uncertain disturbances. To this end, they need to assume the generalized functional safety attributes schemed for comprehensively addressing functional safety problems and cybersecurity problems (including some information security problems) in an integrated way.

5 New security problems brought by generative AI models

The previous two years have seen a lot of progress in generative AI, from the AI models such as DALL-E 2 [36] and Stable Diffusion [37] that have shattered the state-of-the-art of image generation, to the human-like dialogue robots represented by ChatGPT [5], GPT4 [4], Baidu's ERNIE Bot [38]. Generative models have caused a stir on the Internet and have suddenly been on everyone's lips owing to their potential to upend the world of content production. Noteworthily, generative pre-trained AI models represent a transition of AI technologies from perceiving and understanding the world to building and creating the world. They are expected to usher in a new era of AI.

At present, generative pre-trained models have become new content-creation engines. They can create novel content aligned with users' intentions and automatically complete such tasks as text and article production, conversation, translation, coding, painting, video production, *etc.* The ensuing security challenges to the cognitive domain exist mainly in three aspects:

Firstly, generative models lend criminals the convenience to produce disinformation or hot topics generally used in malicious ways. For example, recent advances in controllable text generation techniques such as InstructGPT [39] models may provide favorable conditions for carefully crafted disinformation. Ultimately, there will be more deviations from facts and the network will be flooded with more malicious traffic, thus bringing about "cognitive difficulties".

Secondly, the guidance and feedback mechanisms of generative models make them vulnerable for ill-intentioned individuals or criminals to input false or misleading information to result in outputs that fuel cognitive biases. Once carefully exploited, these biases will breed financial frauds or even support illegal incitements to violence and crime.

Thirdly, generative models such as ChatGPT can automatically generate phishing emails, malicious links, and other techniques of social engineering attacks. When mixed with normal information, it can obfuscate malicious links within normal comments. When users unconsciously click on the hyperlink, it will activate subsequent attack paths and consequently endanger users' personal and property safety.

6 Endogenous safety and security problems brought by AI technologies and new challenges in the cognitive domain

In the AI era, generative AI has made astounding progress and intelligence-driven machines have been involved in the production, dissemination, and feedback chain of information in the cognitive domain. For example, AI anchors have started the "secondary creation" mode of 24-h news broadcasting, intelligent customer service has become the "best salesman" for product consultation and after-sales service, and algorithms for personalized recommendations have built an information feedback ecology based on "human behavior". Machine behavior has become an important emerging variable in cognitive domain analysis. Although existing AI technologies are not yet capable of controlling the automatic setting of topics, they can push the heating and trend of topics through active engagement in topic interactions and have become the important "hidden hands" to detonate the public opinion field at any time. It can be seen that in the era of all-media interconnection, information dissemination in the cognitive domain has represented a new paradigm of "algorithm-led" computational propaganda.

However, due to the limitations of data-driven AI in algorithm explainability and robustness, AI-enabled information dissemination in the cognitive domain also has some defects, which are mainly manifested in three aspects. First, the diversity of text expressions in the open world leads to a low recognition accuracy of cognition-penetrating content. Second, due to the black-box characteristics of AI technologies, algorithm-driven information dissemination tools with regard to the cognitive domain are faced with various security risks such as poisoning attacks, reasoning attacks, and model thefts. Criminals can achieve the goal of confusing and bypassing the supervision mechanism of AI algorithms by constructing texts and images with properties to support cognitive confrontation. Third, the accuracy of existing algorithms, such as intelligent anchoring, intelligent commenting, and intelligent posting, mostly depends on the training corpus and its statistical probability distribution. However, because of the complex and diverse values of the online social corpus, there are often cognitive biases in the content generated by these algorithms, and the corresponding correction job requires cognitive cleaning of a massive corpus, which is difficult to achieve in engineering.

In order to solve the new problems brought forth by AI technologies to the cognitive domain, researchers have proposed approaches to defending AI systems against adversarial attacks, including data sample cleaning and preprocessing, adversarial reinforcement training for algorithms, and multi-model-based integrative defense. These approaches have proved to be effective to some extent in solving the endogenous safety and security problems of the AI algorithm itself. As far as the cognitive domain is concerned, however, AI technologies will also bring new challenges to cognitive security.

Firstly, it is difficult to achieve cyber-cognitive situational awareness. Due to the explosive growth and fission dissemination of information brought about by intelligence technology, conventional cyber-cognitive situation analysis is faced with the dilemma of "too much data and insufficient processing means", leading to the difficulty in realizing an all-domain, full-time and all-round picture of domestic and foreign situations.

Secondly, it is difficult to distinguish malicious social media content. The emergence of new media technologies in large numbers has caused the indiscriminate spreading of junk information, resulting in low credibility and low authority of information. Criminals often arrange and combine a series of topics to not only let the audience "see what you want them to see", but more importantly, let the audience "think the way you expect", to manipulate and guide them to form cognitive deviations from social reality.

Thirdly, cognitive confrontation strategies are stereotyped and passive. Due to the diversity of cognition intervention means and methods, cognitive confrontation strategies also must be accordingly diversified. However, the current adversarial defense strategies for improving cognitive security are relatively rigid for the following reasons. First, it is difficult to find abnormal clues from the explosion of social information. Second, it is difficult to accurately identify malicious cognitive manipulations. Third, it is difficult to choose the most scientific and reasonable adversarial defense strategies for cognitive security. Fourth, it is difficult for current AI technologies to dynamically adapt to the cognitive situation in the process of multiple rounds of cognitive confrontation (see Figure 2).

7 Exploration of mimicry technology in addressing endogenous safety and security problems brought forth by AI to the cognitive domain

The development and application of AI technologies, especially the emergence of large models like ChatGPT and GPT-4, has greatly changed the information dissemination ecosystem in the cognitive domain. On the one hand, it has brought forth the endogenous safety and security problems of AI algorithms. On the other hand, it has also induced new challenges in the cognitive domain. Building more secure and reliable AI algorithms for the cognitive domain is the only way to the transformation of AI applications in this domain. Because current AI technologies are still relatively immature and blunt in terms of cognitive situational awareness, content production, and cognitive guidance, we attempt in the following part to discuss the endogenous safety and security problems brought forth by mimicry technology-based AI systems to the cognitive domain. Then we will go further to propose an endogenous safety and security vision for AI applications.

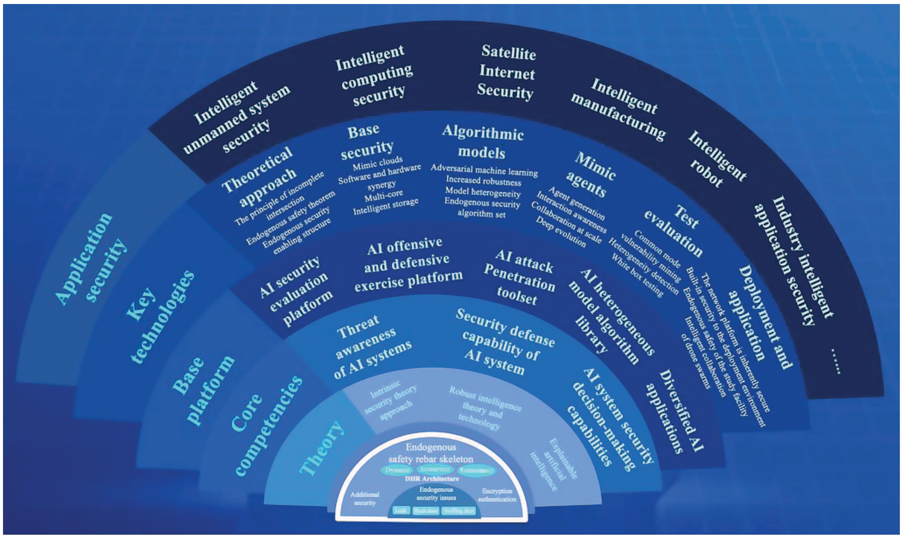
Application security
Intelligent unmanned system security
Intelligent computing security
Satellite Internet Security
Intelligent manufacturing
Intelligent robot
Industry intelligent application security
……
Key technologies
Theoretical approach
The principle of incomplete intersection
Endogenous safety theorem
Endogenous security enabling structure
Base security
Mimic clouds
Software and hardware synergy
Multi-core
Intelligent storage
Algorithmic models
Adversarial machine learning
Increased robustness
Model heterogeneity
Endogenous security algorithm set
Mimic agents
Agent generation
Interaction awareness
Collaboration at scale
Deep evolution
Test evaluation
Common mode vulnerability mining
Heterogeneity detection
White box testing
Deployment and application
The network platform is inherently secure
Built-in security to the deployment environment
Endogenous safety of the study facility
Intelligent collaboration of drone swarms
Base platform
AI security evaluation platform
AI offensive and defensive exercise platform
AI attack Penetration toolset
AI heterogeneous model algorithm library
Diversified AI applications
Core competencies
Threat awareness of AI systems
Security defense capability of AI system
AI system security decision-making capabilities
Theory
Intrinsic security theory approach
Robust intelligence theory and technology
Explainable artificial intelligence
Endogenous safety rebar skeleton
Dynamic
Isomerism
Redundancy
DHR Architecture
Additional security
Endogenous security issues
Encryption authentication
Leak
Back door
Stuffing door

Figure 2. An endogenous safety and security vision for AI applications.

7.1 Mimicry technology and mimicry intelligence

Mimicry refers to an ecological adaptation phenomenon in which one organism mimics another organism in terms of morphology, behavior, and other characteristics, thereby benefiting one or both parties. In 2007, inspired by the mimic octopus in nature, Academician Wu Jiangxing proposed the mimicry computing theory which has realized flexible and reconfigurable computing based on the idea of equivalent functions but different structures. This theory has become the cornerstone of future intelligent computing. In 2013, based on the same idea, Academician Wu put forward another theory of cyberspace mimicry defense, which maintains that the dynamic random adjustment in structures will support higher security gains.

For the scenario in the cognitive domain, behavioral mimicry will help make intelligence-driven machines more human-like in behavior, language, and style. Inspired by the mimicry in biology, we believe that the intelligence needed in the cognitive domain shall be one that can dynamically adapt to the virtual public opinion environment, namely, mimicry intelligence technology. It shall have at least three features. First, it can analyze the information content and traffic mode in its cyber environment, and then generate the content required for interaction. Second, it can interact and collaborate with various entities (machines and humans) in the cognitive domain. Third, it can learn new abilities and dynamically adjust behavioral strategies in response to environmental changes. These three features of mimicry intelligence will help to address both the endogenous safety and security problems of AI systems in the cognitive domain, and the new security challenges in the domain induced by AI.

7.2 Mimicry intelligence framework for cognitive security

Current social media network represents some new features, such as the proliferation of false information, the low threshold of intelligent content generation, and the technological advances in anti-supervision content generation. To solve this problem, we propose a framework of mimicry intelligence for cognitive security. This framework is oriented towards the new environment where "physical, cyber and cognitive" domains are intertwined, especially towards addressing the new problems facing the cognitive domain that can hardly be solved with traditional manual, automatic, and intelligent paradigms. It is based on mimicry intelligence technology and attempts to propose a new idea for addressing security problems in the cognitive domain.

Implementation of the intelligence system framework is mainly concerned with developing a mimicry agent, which refers to a new type of social regulation analyst with the features of "mimicry intelligence", *i.e.*, one with human-like intelligence, able to independently perform diverse tasks, capable of generating social media content similar to human beings, and can collectively complete various cognitive security defense tasks. Specifically, the mimicry agent shall have the following functions.

7.2.1 Highly reliable and highly credible mimicry intelligence guidance

Based on generative AI models, it can, in view of a transmission chain full of cognitive biases, timely introduce mechanisms of intelligent content identification, retweets, comments, and debates with human-like value judgments. Moreover, it can build a system for multiple authentications and interactive feedback of online information resembling the way humans do, thus providing important support for developing robust agents and establishing different communication nodes in the hierarchy.

7.2.2 Localized mimicry migration of diverse knowledge systems

This involves employing intelligent algorithms of data analysis to enable the analysis of massive online data and comprehensive utilization of complex multilingual social and economic knowledge. The task here is to immediately determine value orientation, grasp clues and make topic selections, and employ mechanisms similar to human lifelong learning to realize continuous monitoring and timely correction with the language habits and language styles that users are willing to accept.

7.2.3 Mimicry fusion of multi-source information

This involves the process of employing generative pre-training models and taking together multi-factors such as attention, situation, and value judgment to generate multi-modal content aligned with mainstream values through text generation, speech synthesis, video generation, and other methods. In short, it can provide immediate and high-quality content generation techniques for cyber cognitive security defense.

7.2.4 Mimicry collaboration of multi-agents

This involves applying multi-agent technology to the grouping and collaboration of social robots. For example, in adversarial dialogues, social robots are used to push information to specific groups to provide a “covering”. Moreover, by employing and amplifying the influence of “anthropomorphic” agents, using “focus” agents, follower agents, and roadblock agents to carry out programming and collaborative defense, and meanwhile launching agenda-setting for salient issues or hot topics, purposes such as showing strength and pressure build-up can be achieved.

7.3 Key technologies requiring further breakthroughs

Although generative AI technology has made remarkable breakthroughs and has provided irreplaceable techniques for the application of mimicry intelligence in cognitive security, there are still several crucial technologies that need to solve.

7.3.1 Establishing a theoretical framework for advanced machine cognition from humanoid and brain-like perspectives

Construct mechanisms and methods through which value judgment can be calculated, explained, and predicted based on the cognitive performance of a target object, and explore systematic methods of deep semantic understanding, cognitive recognition, value judgment, and analytical reasoning based on the fusion of multi-source cognitive knowledge [40–42].

7.3.2 Exploring differences in cognitive biases between the digital brain and the biological brain

Analyze the commonalities and differences between the digital brain (computer systems) and the biological brain (cognitive individuals and groups). Study their cognitive “weaknesses” through simulations and psychological experiments to provide the objective basis for cognitive confrontation and effect evaluation [43–45].

7.3.3 Exploring interactions between the cognitive environment and the cognitive subject

Based on the theories and methods of cognitive science, brain science, and psychology, carry out the effective cognitive intervention, and control over target individuals, groups, and important objects, and describe the impact mechanism of the cognitive environment (cyber information environment) on the cognitive subject (humans and intelligent agents) through qualitative and quantitative analyses, to provide theoretical support for the realization of cognitive escape and cognitive immunity [46–48].

7.3.4 Proposing a co-evolution strategy of heterogeneous agent groups

Based on the modeling of the cognitive domain and cognitive representation space, build up heterogeneous (or diverse) multi-agent groups to participate in and influence cognitive activities in a cooperative manner, and in the process, to conduct autonomous learning and offensive and defensive strategy optimization. In a complex environment with multi-party participation, carry out research on influence-oriented strategy generation technology with regard to multi-agent group activities. Under the guidance of a mentor (goal setter), carry out research on collaborative concurrent learning algorithms for heterogeneous multi-agent groups. Study the behavior modeling of other agents and agent groups in the cognitive domain hidden in the online information. Use the global semantic workspace mechanism to explore the realization of an intelligent "brain" composed of multi-agents with "consciousness", and to describe the multi-party offense and defense game of debate through reinforcement learning in the embedded semantic space [49–51].

7.3.5 Developing robust and interpretable AI algorithms

Developing AI models that are resistant to adversarial attacks, biases, and other vulnerabilities. This includes researching robust optimization techniques and creating models that are secure by design. Advanced threat detection and mitigation techniques are required to identify and respond to AI-specific threats, such as adversarial attacks or AI-generated malicious content. AI systems need to be enhanced in their transparency and understandability enable better decision-making and promote trust. We also should develop privacy-enhancing technologies, such as federated learning, differential privacy, and secure multi-party computation, to ensure that AI models can be trained and used without compromising the privacy of users' data. The security of AI infrastructure should be enforced, including cloud platforms and edge devices, to protect against cyberattacks targeting AI systems.

7.3.6 Monitoring, auditing, and regulation

Standardized practices should be established for monitoring and auditing AI systems to ensure compliance with regulations, ethical standards, and best practices. Besides, policy and regulation need to be implemented effectively to address security concerns, promote transparency, and hold AI developers accountable for their systems' performance and behavior. Ethical considerations also should be integrated into AI development and deployment, including the development of ethical guidelines and frameworks for AI systems. Industry-wide security standards and best practices are really required to minimize security risks and improve the overall security posture of AI systems.

7.3.7 Enhancing human-AI collaboration

Effective collaboration between humans and AI systems needs to be fostered to ensure optimal decision-making and enhance the ability to detect and respond to security threats. Enhancing human-AI collaboration involves optimizing communication, building trust, and leveraging the strengths of both humans and AI to create effective teams. We can leverage the complementary strengths of humans and AI systems by identifying the unique skills and abilities of humans and AI systems and assigning tasks accordingly. For example, AI can handle large data processing and pattern recognition, while humans excel at creativity and empathy. During the development of AI systems, we could encourage feedback loops between humans and AI systems to foster mutual learning and improvement. This includes refining AI algorithms based on human input and allowing humans to learn from AI-generated insights.

In addition, we should regularly assess the effectiveness of the collaboration and make adjustments as needed. This may involve refining AI models, updating communication protocols, or redefining roles and responsibilities.

8 Conclusion

In the all-media era, the rapid development of information and cyber technologies represented by big data and AI has boosted the influence of information dissemination in the cognitive domain in a wider, deeper, and stronger way. On the one hand, they provide an important technical approach for purifying the "noise" in the digital space; on the other hand, due to the data security, model security, and framework security concerns of AI systems, the current cognitive domain is also faced with endogenous safety and security problems and new challenges induced by generative AI technology. Therefore, cognitive security needs more robust AI technologies that are better in line with cognitive domain scenarios. In response to this demand, this work proposes a framework for the Mimicry Intelligence technology to address cognitive security problems, which makes full use of the structural effects and adaptability features of the mimicry technology. It is hopeful to promote humanization, localization, and socialization of AI algorithms and applications, and then to provide highly reliable, highly credible, and highly available techniques for cognitive domain games. However, it must be noted that the Mimicry Intelligence framework proposed here also has its own limitations. It is designed for a dynamic and interactive environment. When it comes to an isolated individual, the Mimicry Intelligence framework shall be applied in combination with other strategies and techniques in order to achieve the desired effects.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

No data are associated with this article.

Authors' Contributions

Ruiyang Huang contributed to the study conception and manuscript drafting. Xiao Qing jointly wrote this manuscript and revised it critically for important intellectual content. Yuting Shang contributed to language polishing and content revision. Xiangyang Xue supervised the study, developed the theoretical framework, and contributed to content revision. All authors reviewed the results and approved the final version of the manuscript.

Acknowledgements

We are very grateful to all editors and reviewers for their constructive comments. In addition, we thank Hong Zou for revising and polishing the content of the manuscript.

Funding

This work was supported in part by National Key R&D Plan (2022YFB3102901).

References

[1] Brown T, Mann B and Ryder N et al. Language models are few-shot learners. Adv Neural Inf Process Syst 2020; **33**: 1877–901.
[2] Chowdhery A, Narang S and Devlin J et al. Palm: Scaling language modeling with pathways, arXiv preprint arXiv:2204.02311, 2022.
[3] OpenAI. ChatGPT: Optimizing language models for dialogue. OpenAI Blog, 2022.
[4] OpenAI. GPT-4 technical report. OpenAI, 2023.
[5] Guo B, Ding Y and Sun Y et al. The mass, fake news, and cognition security. Front Comput Sci 2021; **15**: 1–13.
[6] Andrade RO and Yoo SG. Cognitive security: A comprehensive study of cognitive science in cybersecurity. J Inf Secur Appl 2019; **48**: 102352.
[7] Jobson KO and Hartley DS. Achieving cognitive warfare superiority amidst accelerating change. Phalanx 2022; **55**: 28–31.
[8] Claverie B, Prébot B, Buchler N, and Du Cluzel F. Cognitive Warfare: The Future of Cognitive Dominance, NATO Collaboration Support Office, 2022; **2**: 1–7.
[9] Chen YF, Ye H and Wang DD. Theories of social preferences beyond homo economicus: A review based on the experimental economics. Nankai Econ Stud 2012; **01**: 63–100.
[10] Zhou LN, Yang Z and Chu BL, et al. Overview of multimedia cognition security. J Signal Process. 2021; **37**: 2440–2456. https://doi.org/10.16798/j.issn.1003-0530.2021.12.012

[11] Fan WJ and Wang YB. Cognition security protection about the mass: A survey of key technologies. J Commun Univ China Sci Technol 2022; **29**: 1–8.

[12] Toffler A. The third wave: The classic study of tomorrow. Bantam, 2022.

[13] Ke P, Zou JH and Sun XN. Launch a new round of strategy for the transformation and upgrading of cultural industry: Analysis and inspiration of "opinions on promoting the implementation of the national cultural digitalization strategy". Inf Stud: Theory Appl 2022; **45**: 1.

[14] Li Y, Li X, Shen S, et al. DTBVis: An interactive visual comparison system for digital twin brain and human brain[J]. Visual Informatics, 2023, ISSN 2468-502X, https://doi.org/10.1016/j.visinf.2023.02.002.

[15] Wang WX, Zhou F and Wan YL et al. A survey of metaverse technology. Chin J Eng 2022; **44**: 744–56.

[16] Parthasarathy PK, Mantri A and Mittal A et al. Digital brain building a key to improve cognitive functions by an EEG–controlled videogames as interactive learning platform. In: Congress on Intelligent Systems: Proceedings of CIS 2020. Singapore: Springer, 2021, vol. 1, 241–52.

[17] D'Angelo E and Jirsa V. The quest for multiscale brain modeling[J]. Trends Neurosci, 2022; Oct, **45**: 777–790. https://doi.org/10.1016/j.tins.2022.06.007. Epub 2022 Jul 27. PMID: 35906100.

[18] Avramovic P, Rietdijk R and Attard M et al. Cognitive and behavioral digital health interventions for people with traumatic brain injury and their caregivers: A systematic review. J Neurotrauma 2023; **40**: 159–94.

[19] Roetzer-Pejrimovsky T, Moser AC and Atli B et al. The digital brain tumour atlas, an open histopathology resource. Sci Data 2022; **9**: 55.

[20] O'Sullivan ED and Schofield SJ. Cognitive bias in clinical medicine. J R Coll Physicians Edinb 2018; **48**: 225–32.

[21] MacLeod C and Mathews A. Cognitive bias modification approaches to anxiety. Annu Rev Clin Psychol 2012; **8**: 189–217.

[22] Acciarini C, Brunetta F and Boccardelli P. Cognitive biases and decision-making strategies in times of change: A systematic literature review. Manag Decis 2021; **59**: 638–52.

[23] Binz M and Schulz E. Using cognitive psychology to understand GPT-3. Proc Natl Acad Sci 2023; **120**: e2218523120.

[24] Szegedy C, Zaremba W and Sutskever I et al. Intriguing properties of neural networks. In: 2nd International Conference on Learning Representations, ICLR 2014, 2014.

[25] Ian JG, Jonathon S and Christian S. Explaining and harnessing adversarial examples. In: Proceedings of the International Conference on Learning Representations, 2015.

[26] Jia R and Liang P. Adversarial examples for evaluating reading comprehension systems. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, 2021–2031.

[27] Li Y, Lyu X and Koren N et al. Anti-backdoor learning: Training clean models on poisoned data. Adv Neural Inf Process Syst 2021; **34**: 14900–12.

[28] Li Y, Jiang Y and Li Z et al. Backdoor learning: A survey. In: IEEE Transactions on Neural Networks and Learning Systems, 2022.

[29] Lyu L, Yu H and Ma X et al. Privacy and robustness in federated learning: Attacks and defenses. In: IEEE Transactions on Neural Networks and Learning Systems, 2022.

[30] Xiao H, Biggio B and Brown G et al. Is feature selection secure against training data poisoning? In: International Conference on Machine Learning. PMLR, 2015, 1689–98.

[31] Zhengli Z, Dheeru D and Sameer S. Generating natural adversarial examples. In: Proceedings of the International Conference on Learning Representations, 2018.

[32] Yuan X, He P and Zhu Q et al. Adversarial examples: Attacks and defenses for deep learning. IEEE Trans Neural Netw Learn Syst 2019; **30**: 2805–24.

[33] Saha A, Subramanya A and Pirsiavash H. Hidden trigger backdoor attacks. Proc AAAI Conf Artif Intell 2020; **34**: 11957–65.

[34] Minhao C, Wei W and Cho-Jui H. Evaluating and enhancing the robustness of dialogue systems: A case study on a negotiation agent. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.

[35] Quiring E and Rieck K. Backdooring and poisoning neural networks with image-scaling attacks. In: 2020 IEEE Security and Privacy Workshops (SPW). IEEE, 2020, 41–7.

[36] Kahn J. Move over photoshop: OpenAI has just revolutionized digital image making. Fortune, 2022.

[37] Rombach R, Blattmann A and Lorenz D et al. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 10684–95.

[38] Baidu: Baidu to launch ERNIE Bot-integrated cloud service on March 27. [2023-03-21], https://finance.sina.cn/2023-03-21/detail-imymqrre3128262.d.html

[39] Ouyang L, Wu J and Jiang X et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst 2022; **35**: 27730–44.

[40] Fei N, Lu Z and Gao Y et al. Towards artificial general intelligence via a multimodal foundation model. Nat Commun 2022; **13**: 3094.

[41] Ororbia A and Kifer D. The neural coding framework for learning generative models. Nat Commun 2022; **13**: 2064.

[42] Le Cun Y. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. Open Review, 62. Corpus ID: 251881108 Computer Science. https://openreview.net/pdf?id=BZ5a1r-kVsf.

[43] Williamson B. Brain data: Scanning, scraping and sculpting the plastic learning brain through neurotechnology. Postdigital Sci Edu 2019; **1**: 65–86.

[44] Mehonic A and Kenyon AJ. Brain-inspired computing needs a master plan. Nature 2022; **604**: 255–60.

[45] Ji X, Dong Z and Lai CS et al. A brain-inspired in-memory computing system for neuronal communication via memristive circuits. IEEE Commun Mag 2022; **60**: 100–6.

[46] Szabo B, Valencia-Aguilar A and Damas-Moreira I et al. Wild cognition–linking form and function of cognitive abilities within a natural context. Curr Opin Behav Sci 2022; **44**: 101115.

[47] Hutchins E. The distributed cognition perspective on human interaction. In: Roots of Human Sociality. Routledge, 2020, 375–98.

[48] Magnani L and Magnani L. AlphaGo, locked strategies, and eco-cognitive openness. Eco-Cognit Comput: Cognit Domest Ignorant Entities 2022; **43**: 45–71.

[49] Zhao Z and Zhang X. A continuous heterogeneous-agent model for the co-evolution of asset price and wealth distribution in financial market. Chaos Solitons Fractals 2022; **155**: 111543.

[50] Fu Q, Ai X and Yi J et al. Learning heterogeneous agent cooperation via multiagent league training, arXiv preprint arXiv:2211.11616, 2022.

[51] Rizk Y, Awad M and Tunstel EW. Cooperative heterogeneous multi-robot systems: A survey. ACM Comput Surv (CSUR) 2019; **52**: 1–31.

Ruiyang Huang is currently an associate research fellow of cybersecurity at the National Digital Switching System Engineering & Technological R&D Center (NDSC), Henan, China. His research interests include information content security, cognitive security, and artificial intelligence.

Xiaoqing Zheng is an associate professor at the School of Computer Science at Fudan University, Shanghai, China. He received his Ph.D. degree in computer science from Zhejiang University in 2007. He has been doing research on semantic data integration during his stay at the information technology group, Massachusetts Institute of Technology (MIT) as an international faculty fellow. He also visited the natural language processing and machine learning groups at the University of California, Los Angeles (UCLA) as a visiting researcher. His research interests include natural language processing and machine learning.

Yuting Shang is a research assistant at the Institute of Big Data at Fudan University, Shanghai, China. With a MA degree in English Language and Literature, her research interests include social engineering and cognitive security.

Xiangyang Xue received B.S., M.S., and Ph.D. degrees in communication engineering from Xi'dian University, Xi'an, China, in 1989, 1992, and 1995, respectively. He is currently a professor of computer science at Fudan University, Shanghai, China. His research interests include multimedia information processing, cognitive security, and machine learning.