DOI 10.15302/J-SSCAE-2021.03.004

Development of Content Security Based on Artificial Intelligence

Zhu Shiqiang, Wang Yongheng

Zhejiang Lab, Hangzhou 311121, China

Abstract: Content security refers to the protection of information content that meets requirements at the political, legal, and moral levels. Recent developments in artificial intelligence (AI) have significantly impacted content security. In this article, we summarize the research status and development trends of AI-based content security in China and abroad based on the major strategic demands and present the key technical issues regarding AI-based content security. This study proposes building the world's leading AI-based content security system through a three-step strategy. Innovation and breakthroughs should be made in areas such as adversarial machine learning, explainable AI, hybrid enhanced intelligence, and knowledge-driven content security. Additionally, the establishment of policies, regulations, and regulatory mechanisms should be emphasized. Furthermore, major content security infrastructure such as cyber ranges for content attack and defense and large-scale social system simulation devices for public opinion attack and defense should be established.

Keywords: artificial intelligence (AI); content security; system construction

1 Introduction

With the rapid development of mobile Internet, digital media, artificial intelligence (AI), and other related technologies and applications, content security has become critical. Content security has two main aspects. The first refers to the protection of information content (generally using methods such as anti-theft and anti-tampering), which involves ensuring information content confidentiality, intellectual property protection, information hiding, and privacy protection. The second aspect refers to making sure the content meets the requirements of the political, legal, and moral standards of the country. As far as China is concerned, the information content should be politically healthy and conform to the country's laws and regulations and to the ethical standards of the nation. Content security is important for any country and its society from many perspectives: highly reactionary content can jeopardize national security, risky content such as pornography and gambling-related content can have adverse impact on people's livelihoods, and content such as spam can affect businesses as well as personal lives. Furthermore, content fraud, including content such as fake news and online rumors, has nowadays become one of the most significant issues in content security governance. From the perspective of information review, digital information can exist in large quantities, have diversified sources and highly complex content, and can be disseminated rapidly; hence, the approach used to fact-check it needs to be highly efficient, rapid, and accurate. In the early days of COVID-19, fake news and online rumors posed serious obstacles to the prevention and control of the pandemic. Furthermore, with the current popularity of live webcasting, some individuals and organizations have begun to use live broadcasts to conduct various illegal and criminal activities. Therefore, real-time content supervision of massive live broadcast data is critical.

In recent years, AI has been widely used in the field of content security, achieving many breakthroughs.

Received date: January 11, 2021**; Revised date:** February 20, 2021

Corresponding author: Zhu Shiqiang, professor of Zhejiang Lab. Major research field is robot and artificial intelligence. E-mail: zhusq@zhejianglab.com

Funding program: CAE Advisory Project "Strategic Research on the Secure and Independent Development of New-Generation Artificial Intelligence" (2019-ZD-01)

Chinese version: Strategic Study of CAE 2021, 23(3): 067–074

Cited item: Zhu Shiqiang et al. Development of Content Security Based on Artificial Intelligence. *Strategic Study of CAE,* https://doi.org/10.15302/J-SSCAE-2021.03.004

However, content security algorithms based on AI may suffer from data sample contamination and adversarial algorithm attacks, resulting in serious decision-making errors. Furthermore, content deception technologies, such as forged images, fake news, and voice fraud based on deep learning, can be used to disseminate fake information. Criminals have used intelligent recommendation algorithms to make the dissemination of fake information more targeted and concealed. However, more recent advancements made in AI have provided new opportunities for ensuring content security. Technologies such as deep learning and knowledge graphs can effectively improve content identification, protection, and violation review capabilities, and accelerate the automation, intelligence, efficiency, and precision of content security governance.

In March 2019, the Chinese Academy of Engineering launched a major consulting project, “Strategic Research on the Secure and Independent Development of New-Generation Artificial Intelligence.” This paper serves as an academic display of the research results of the project “Research on Content Security and Attack and Defense Strategy Based on Artificial Intelligence.” Starting from the major strategic needs of AI-based content security, it sorts relevant key technologies and applications that can be used for content security, summarizes the research status and development trends in China and abroad, and proposes recommendations for AI-based content security development in China.

2 Key technologies and applications of content security based on AI

2.1 Key technologies

2.1.1 AI-based content forgery and protection

The development of AI, especially deep learning, has made content forgery more convenient. Deep fake uses AI programs and deep learning algorithms to realize visual and audio simulation and forgery [1,2]. The technologies involved in deep forgery mainly include autoencoders [3] and generative countermeasure networks [4]. Presently, the deep forgery technology not only changes the face but also simulates the voices of real people and creates portraits of people that do not exist in reality. Combined with AI-based natural language generation technology and social network communication, deep forgery has played a major role in the evolution of fake news [5]. This in-depth forgery technology with digital automation features uses various media to disseminate false information. It has strong dissemination potential and can achieve large-scale and latent political manipulation and control, thus significantly aggravating political security threats from the cyberspace and the complexity of the countermeasures to fight them [6].

To tackle the menace created by content forgery technologies, a large number of false content detection technologies has emerged. Feature extraction, an AI-based false content detection method, mainly includes the GAN pipeline, deep learning, and steganalysis. The classifiers mainly include support vector machines (SVMs) and convolutional neural networks (CNNs) [7]. Furthermore, multiple AI-based methods, involving deep learning, knowledge base, graph data mining, and many other related technologies, have been proposed for detecting false news, generally using knowledge base, writing style, communication characteristics, birthplace, etc. [8]. Nevertheless, detection of new types of false news, early detection of false news, cross-domain detection, and interpretable detection still suffer from significant challenges.

2.1.2 Content-analysis-oriented AI model and algorithm security

AI-based content analysis involves various machine learning models and algorithms for text, image, and audiovisual processing. The security of these models and algorithms has a significant impact on content security. The security of machine learning models and algorithms mainly involves the following aspects [9]:

(1) Poisoning attack and defense. A poisoning attack is a type of induced attack that damages the usability and integrity of the model, mainly contaminating the required training data while the model is being trained. The attacker destroys the probability distribution of the original training data by injecting carefully forged malicious data samples.

(2) Backdoor attacks and defense. Backdoor attacks are used to implant backdoors in neural network models through two ways: data and models. When obtaining a specific input, the model is triggered and then causes the neural network to produce an incorrect output; thus, these types of attacks are concealed and difficult to discover.

(3) Adversarial attacks and defense. Machine learning and neural network models are vulnerable to adversarial attacks. In such attacks, by adding specific perturbations to the original samples, the classification model can make incorrect classification judgments for the newly constructed samples.

(4) Model extraction and defense. This refers to stealing models or restoring training data members through

DOI 10.15302/J-SSCAE-2021.03.004

black box detection, such as stealing stock market prediction models and spam filtering models.

Presently, a large number of new machine learning algorithms appear every year, and the security of these algorithms has become a widely shared concern. Model and algorithm security issues can be regarded as a tug of war between the defense and the attacker for modeling each other in the absence of information. Therefore, in-depth research is necessary on new robust models and training algorithms, multilearner security issues, and system modeling and reasoning issues in the absence of information.

2.1.3 Content-analysis-oriented interpretive AI

AI technology, represented by deep learning, faces the problem of interpretability. When it is applied to content analysis in sensitive fields, it is difficult for the “black box” algorithm that lacks transparency and intelligibility to gain people’s sense of security and trust.

There are three main directions in the interpretability research of AI models [10]: (1) Deep interpretation, that is, the use of new deep learning models to learn features that can be used for interpretation. Many related works have been combined with visualization technology to provide a more intuitive explanation on this front. (2) Explainable model. The traditional Bayesian, decision tree, and other models have good interpretability. Presently, many researchers have improved the deep learning model to make it more interpretable. (3) Model reasoning. This method regards the machine learning model as a black box and builds a new interpretable model externally through a large number of experiments. A new research method involves building a set of machine learning techniques that can automatically generate interpretable models and maintain high learning efficiency.

Although research on model interpretability has achieved some remarkable results, it is still in its infancy and is up against multiple challenges that need to be addressed. One of the current challenges is the design of more accurate and friendly interpretation methods to eliminate the inconsistency between the interpretation results and true behavior of the model. Another challenge is to design more scientific and uniform interpretability evaluation indicators so that performance and safety of interpretable methods can verified [11].

2.2 Applications

2.2.1 Network public opinion analysis and supervision

Big data and AI technologies have provided us with new resources, methods, and paradigms for the analysis, research, and evaluation of public opinion [12]. Public opinion is a complicated piece of content and requires instant analysis. It has been shown that AI can make public opinion analysis more efficient and accurate, and in turn reduce labor costs.

In recent years, AI-based network public opinion analysis and supervision have been widely used. Baidu’s media public opinion analysis tools are oriented to traditional and new media industries and provide public opinion analysis capabilities for content production, opinion and transmission analysis, operational data display, and other business scenarios. Government affairs public opinion analysis tools rely on web content mining capabilities and Chinese semantic analysis technology to support the in-depth exploration of domestic and foreign risk information and for the real-time perception of urban public opinion. Tencent’s WeTest public opinion monitoring tool uses distributed crawlers to capture the mainstream application market (e.g., AppBao) reviews, star ratings, and user post discussions on mainstream forums (Baidu Tieba, etc.), and intelligently aggregates user comments and classifies them. Through sentiment analysis and sentiment dimension extraction technology, intelligent analysis and the positioning of specific problems can be achieved. In 2020, the AI Industry Development Alliance launched by the China Academy of Information and Communications Technology (CAICT) and other units released a report that pointed out that AI and big data played a significant role in the analysis of public opinion regarding the COVID-19 epidemic.

2.2.2 Multimedia content analysis and review

Even in countries that promote freedom, such as Norway, Japan, and Italy, Internet content censorship is intensifying. For example, A team at the University of Michigan used its Censored Planet tool (an automatic censorship tracking system launched in 2018) to collect more than 21 billion content review measurements from 221 countries in the past 20 months [13]. Recently, multimedia data, especially video data, have seen unprecedented growth and it is expected that this trend will continue to grow. Today, massive multimedia data far exceed human processing capabilities, and therefore, AI has been widely relied upon for content analysis and reviews.

Multimedia content analysis based on AI mainly includes intelligent review, content understanding, copyright protection, and intelligent editing. Content review involves the evaluation and identification of unhealthy and potentially harmful content on the Internet, including pornography, violence and terrorism-related content, advertising QR codes, and meaningless and socially sensitive live broadcasts. The content understanding function includes content classification, labeling, character recognition, and voice recognition, as well as the recognition of text in images and videos. Copyright protection functions include content similarity identification, same-source content retrieval, and audio and video fingerprinting. Intelligent editing realizes the generation of video first images, video summaries, and video highlights, and supports news disassembly.

Currently, short videos and pictures have become the main content of multimedia reviews. A range of prohibited multimedia content, such as pornography and content related to terrorism, can be accurately identified through tagging of data and deep learning algorithms. The Artificial Intelligence Rumor Smasher, launched in 2019 by Alibaba, realizes intelligent credibility recognition of news content by analyzing user portraits and matching and verifying them with authoritative knowledge bases in the knowledge map. The accuracy rate for specific scenes was 81%. Based on the accumulated standard sample database, CAICT carried out modeling training for the identification of obscene and pornographic, terrorism, violence, and other illegal information, initially achieving a recognition accuracy rate of over 97%, which is 17% higher than that of the traditional method, and a recognition speed that was 110 times better than that of the traditional method. In February 2021, Baidu released the *2020 Annual Report on Comprehensive Management of Information Security*. The Baidu Content Security Center mined over 51.54 billion pieces of harmful information in 2020 using AI technology, and cracked-down on over 80 million pieces of relevant harmful information through manual and autonomous inspections. The review speed has been greatly increased, and six review dimensions have been formulated, including terrorism, political sensitivity, watermarks, tags, public figures, and malicious images.

3 Development status

3.1 Outside China

3.1.1 The United States

As the most developed country in the content industry, the United States has taken the following offensive as well as defensive steps for ensuring content security: First, in response to the increasingly severe geopolitical threats and domestic ideological security needs, the United States has increased the supervision of the Internet content industry, especially regarding discrimination, prejudice, and other content; second, attaching great importance to the application of AI algorithms in content security, Google and other companies cooperate closely with the government, and the government puts forward relevant review requirements for the security of the algorithms; third, the Senate and House of Representatives of the United States attach great importance to the use of AI algorithms for detecting content fraud, and have incorporated content fraud into the legislative process by holding hearings or proposing related bills.

In 2019, the US government launched the *American AI Initiative*, emphasizing the importance of AI in traditional security and using AI to ensure the position of the United States as the global leader is maintained in response to challenges from “strategic competitors and foreign opponents.” In March 2021, the US National Security Council for Artificial Intelligence issued a final report to actively maintain the dominance of the United States in the field of AI. The report discusses how the United States has won the initiative and maintained its global leadership in the era of fierce competition in AI, and elaborates on the course of action for the future reform of federal agencies. The US Defense Advanced Research Projects Agency (DARPA) launched the “Maven Project” in 2017 to integrate machine learning algorithms in the process of intelligence collection. DARPA has also promoted other research related to AI content security, including media forensics and interpretable AI. In terms of AI-based content security technology, the United States ranks first in terms of citation influence, number of patents, number of companies, and financing scale.

3.1.2 Europe

Europe has paid greater attention to ethics in AI security. The European Commission issued the *Ethics Guidelines for Trustworthy AI* in April 2019, proposing a framework for realizing the full lifecycle of trustworthy AI. The framework proposes seven key elements for achieving trustworthy AI: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination, and fairness; societal and environmental well-being; and accountability.

The guidelines particularly emphasize privacy and data governance, that is, AI must ensure privacy and data

DOI 10.15302/J-SSCAE-2021.03.004

protection throughout the lifecycle of a system. The datasets and processes used by the AI system to make decisions should be recorded for traceability, and transparency should be increased, including around the collected data and data tags used by the algorithm. When the AI system has a significant impact on humans, the AI system can reasonably explain the decision-making process while avoiding unfair discrimination. The United Kingdom, France, Germany, Russia, and other European countries have issued programmatic documents on AI security, which all involve strategies and development plans for AI content security.

3.1.3 Japan

Games, animation, and music are seen as the most developed Internet-based industries in Japan. Therefore, Japan's content security strategy is mainly aimed at the ownership of intellectual property rights of the content produced by AI, and Japan aims to achieve this by bringing in relevant bills. In December 2018, the Cabinet Office released *Human-centered Artificial Intelligence Society Principles*, a highest-level policy document to promote the development of AI in Japan. It demonstrates the position of the Japanese government vis-à-vis the development of industrial intelligence from a macro and ethical perspective.

In the part related to AI content security, the report mainly emphasizes the protection of privacy and guarantee of security. In terms of privacy protection, AI can infer political positions, financial situations, interests, and hobbies of individuals with high accuracy based on their actions and other data. In terms of safety protection, the report emphasizes that while realizing the automation of social systems and improving safety, AI needs to grasp the balance between benefits and risks to improve the safety and sustainability of society as a whole, promote research regarding AI risk and risk reduction, and attach importance to the sustainability of the use of AI.

3.1.4 Others

In addition to the United States, Europe, Japan, and other AI developed countries, many other countries have also launched national strategies for AI and content security. In June 2018, India issued the *Artificial Intelligence National Strategy*, emphasizing how it can use AI as a transformative technology to promote economic growth, increase social inclusion, and seek an AI strategic deployment suitable for developing countries, which can be replicated and expanded in other developing countries. Israel revealed its national AI plan in November 2019, highlighting how AI can be used to meet the needs of the government and military, and its security. The main goal is to ensure that the education system and academic institutions can provide enough AI engineers to meet the manpower needed by the government, national defense military, and industry. The Canadian government, meanwhile, has invested billions of dollars in AI research and development, including the establishment of specialized research institutions and the training of thousands of researchers and computer scientists, with an aim of forming an extremely rich AI ecosystem. Researchers have conducted extensive research on content-security-oriented multimedia analysis, interpretable AI, and other technologies.

3.2 Inside China

3.2.1 National strategy and technology research

China attaches great importance to the development of the AI industry. In 2017, the State Council issued the *Development Plan of New-Generation Artificial Intelligence* as the core driving force of a new round of strategic industrial transformation to lead future development. According to the *New Generation Artificial Intelligence Development Plan 2019* report, China has published the largest number of AI papers in the world, and the number of enterprises and the financing scale rank second in the world. In January 2018, China held the founding meeting of the National Artificial Intelligence Standardization Group and the Expert Advisory Group and released the *Artificial Intelligence Standardization White Paper 2018 Edition* to establish a unified and complete standards system. In 2019, China established the Artificial Intelligence and Security Special Committee of the Chinese Society of Artificial Intelligence, which focused on finding new ways to solve the problems facing the cyberspace security in the country. In addition, in light of the most critical needs of national network information security, the Chinese Academy of Sciences established the National Engineering Laboratory of Information Content Security Technology to carry out research on core key technologies, such as network information acquisition, analysis, and mining.

3.2.2 Innovative development of emerging industries

In terms of content security and other AI-related policies, currently, the local governments in China closely follow the trends adopted at the national level. In line with the decisions made at the central level, local governments have put in place various AI development action plans, strengthened AI supervision, and promoted the wide application of AI in content security.

SenseParrots, a deep learning platform independently developed by SenseTime, has been applied in applications such as face recognition, image recognition, video analysis, unmanned driving, and medical image recognition, providing technical support for AI-based content security. Tencent pays attention to the building of network security capabilities and has established seven network security laboratories, focusing on security technology research and the establishment of security attacks and defense systems. The DeepEye recognition technology engine of Tencent Youtu can be used to analyze content confidence. It relies on the massive sample advantage of Tencent Social to conduct deep recognition training and perform text recognition based on multimodel matching technology to aid content security. The Alibaba Cloud platform uses natural language understanding algorithms to identify text spam and malicious behaviors, and using deep learning algorithms combined with unique intelligence, public opinion, and early warning and analysis systems and real-time updated sample libraries, it quickly locates sensitive information. Huawei has done a great deal of work on the security of machine learning algorithms and has provided solutions to model and algorithm attacks such as data poisoning, model extraction, and backdoor attacks, and provided security for content analysis models and algorithms based on AI. Additionally, many innovative Chinese enterprises have carried out research on applying AI for content security.

3.2.3 New developments in the age of AI 2.0

Pan Yunhe, a member of the Chinese Academy of Engineering, declared the present time as the era of AI 2.0. Its core concepts include expanding, managing, and reorganizing human knowledge through big data and swarm intelligence, providing suggestions for economic and social development, and reaching or even surpassing human capabilities in more specialized fields such as games, recognition, control, and prediction [14]. AI 2.0 is closely integrated with human–computer interaction and is combined with big data and cloud computing. Big data and cloud computing are important driving forces in AI development. AI 2.0 also features the close integration of AI and intelligent surveillance and a combination of AI and advanced manufacturing.

Big data intelligence and cross-media intelligence in the AI 2.0 era will have a significant impact on content security. Based on the deep integration of multimodal data, knowledge base, and cross-media analysis and reasoning with brain-like computing, swarm intelligence, and other technologies, smarter and more accurate content analysis can be realized. Technology companies in China, such as Alibaba, Tencent, Baidu, and Huawei, have begun to actively explore technologies related to AI 2.0 and their applications in content security governance. For example, the special action of Network Ecological Governance 2019 exhibited by the Baidu Content Security Center integrates many AI technologies, such as natural language screening, audio and video intelligent recognition, and content intelligent mining.

4 Strategic suggestions on the development of content security based on AI in China

4.1 Overall development strategy

In general, a three-step development strategy will be adopted: the plan for the development of content security based on AI will achieve initial results by 2025, world-class results by 2035, and world-leading results by 2050. Below we discuss these strategies in detail.

By 2025, the development environment and infrastructure of AI content security will be improved, and key frontier theories and application technologies centered around content security will make remarkable progress. Preliminary results will be achieved in the research of AI models and algorithms for content security, key breakthroughs will be made in the research of key AI technologies for content attack and defense, and a number of efficient AI content security enterprises will emerge. A group of leading talents and experts in the field of security will be gathered, and a three-level AI content security system for individuals, enterprises, and countries will be put in place.

By 2035, the developmental environment of AI content security will have obvious advantages: it will invest heavily in infrastructure development, its scale and level will rank first in the world, and world-class theoretical research on content security AI models and algorithms and the key technologies of content attack and defense will be conducted in the country; furthermore, highly innovative theories and content security methods will be put forward, and the audit mechanism of AI content security will be formed.

By 2050, led by a teams of top international experts and entrepreneurial enterprises in AI-based content security, China will emerge as the leader in the field of AI content security, in terms of the overall innovation ability and the technical application of AI algorithms and technologies for content security. Furthermore, the system of AI content security regulations, ethics, and policies will be established, and the country will achieve world dominance in its ability to assess and control AI security.

4.2 Content security development policy guarantee

4.2.1 Government-led AI development route

The state plays a critical role in maintaining China's network information security, as it enjoys unparalleled power in regulating Internet behavior and combating cybercrimes. The government should lead the development strategy of AI security, raise AI security to the national level, and regard AI security as the core competitiveness of the country's future development, and at the same time regard content security as an important part of AI security. It is necessary to conduct an in-depth analysis of content security requirements in the application of AI, strengthen the top-level design, propose an overall plan based on national cyberspace security, establish implementation rules for content security governance, and establish a security access system and detection and evaluation methods and mechanisms. To promote the development of AI content security with enterprises as the main body, the government should provide guarantees in terms of laws and regulations, security risks, policy guidance, resource allocation, and industry standards; formulate phased development strategies with clear objectives; and strengthen guidance and execution in terms of scientific research projects, intelligent economy, and the overall layout of the intelligent society.

4.2.2 Establishment of a legal and effective supervision mechanism

The AI content security risk management control system should be formulated to protect AI content security from multiple levels including system security, algorithm security, and application security, and security protection measures should be formulated to ensure user data security, avoid the harm caused by the algorithm design to the public, clarify algorithm motivation and interpretability, and overcome unfair effects caused by the algorithm design and data collection. The content security risk management control system provides detailed regulations for the content security of key applications such as social networks, short videos, and online live broadcasts. The security and unity of the target functions and technology can be ensured through the establishment of a reviewable, traceable, and deducible supervisory mechanism. An AI data security supervision mechanism should be established. In accordance with national laws and regulations, government departments carry out supervision and inspection through online and offline methods to detect and prevent security risks in time, aiming at preventing data security risks such as excessive data collection, data bias and discrimination, and abuse of data resources.

4.2.3 Development of AI content security standards system

The establishment of AI content security standardization organizations in China should be optimized, and the joint and orderly promotion of AI content security standards by national, industrial, and group standardization organizations should be promoted. Content security testing and evaluation methods and index systems should be developed for AI products, applications, and services to strengthen content security and privacy protection through testing and evaluation. According to the classification of the content security bearing mode, a security indicator system for graphics/image content, text content, video content, and audio content should be established and classified according to the content security behavior mode. This involves intelligent pornography identification, violence and terrorism-related content recognition, sensitive face recognition, bad scene identification, advertising identification and filtering, logo identification, anti-spam, and other security indicator systems.

4.3 Development of AI technology innovation for content security

In the current socio-political and technological landscape, content security is facing enormous challenges and requires innovative breakthroughs at the technical level, mainly including the following.

4.3.1 Hybrid augmented intelligence based on human–computer cooperation

In terms of content security, the current AI technologies cannot complete tasks independently, such as identifying illegal activities in live videos. Therefore, it is necessary to make breakthroughs in human–computer collaboration, brain–computer collaboration, cognitive computing, and other technologies; fully integrate human and machine intelligence; realize the further enhancement of AI; and realize the continuous improvement of AI technologies based on human guidance and feedback.

4.3.2 Knowledge-driven content security

Complex applications of AI-based content security require the assistance of knowledge; thus, it is necessary to vigorously promote knowledge-driven content security innovation. Technical directions include cross-media knowledge acquisition, content security knowledge base construction, large-scale knowledge base management and knowledge evolution, and content-security-oriented knowledge reasoning.

4.3.3 High-performance content security analysis

Once harmful content is disseminated, it may cause significant losses to the country and society; thus, many content-monitoring applications should emphasize instantaneity. High-performance content analysis algorithms need to be studied, especially in the context of live videos, which need to deal with massive video data streams and correlate multichannel historical data and knowledge.

4.3.4 Adversarial machine learning

Adversarial machine learning directly affects the security of the AI model and algorithm, thus directly threatening content security. It is necessary to make technological innovations to defend against data poisoning and decision-time attacks, and increase the robustness of in-depth learning models and algorithms.

4.3.5 Explainable AI

The interpretability of AI models and algorithms directly affects the credibility of content security analysis and regulatory applications. Technical innovations are, thus, necessary in interpretable machine learning models, model interpretation based on deep learning and visualization, and model interpretation based on reasoning.

4.4 Improvement of content security infrastructure

To promote the development of content security based on AI through technological innovation, it is necessary to establish and improve a number of major national infrastructures to meet the needs of new technology experiments, as well as the need for regulatory policy and strategy evaluation.

4.4.1 Cyber range for content-oriented attack–defense drill and research

A large-scale, open, shared, and continuously growing national content security cyber range should be built to provide users with high-end services such as content security attacks and defense system verification, application system and security product security testing, risk assessment, and emergency response. Building such a cyber range will involve innovating on key and difficult technologies and tasks such as advanced simulation of complex network attributes, behaviors, and interactive dynamics, large-scale simulation of complex business and node reconstruction, antagonistic simulation of panoramic capture reproduction and stress countermeasures, and simulation evaluation of multilevel and multidimensional attack effectiveness. It will also involve constructing an offensive and defensive exercise model in terms of technical verification, strategic prediction, content inspection, intelligence analysis, and public opinion early warning, and substantially improved national content security capabilities through systematic, basic, and pioneering work.

4.4.2 Large-scale social system simulation device for public opinion attack–defense drill and research

This simulation device can be constructed by combining virtual and real data; and real data can be used to drive the virtual model to analyze the virtual and real data in an integrated manner. Based on the latest AI technology, an intelligent fitting model should be established to realize an interactive visual analysis of large-scale public opinion attacks and defense simulations. It should ideally support multiple users in carrying out experimental analysis on the simulation deduction platform, support real-time intervention in the operation of the simulation system, and provide visual data to show the verification effect. Furthermore, the government should have a more comprehensive understanding of public opinion information and dynamics, and should capture potentially sensitive changes in a timely manner.

References

[1] Kietzmann J, Lee L W, McCarthy I P, et al. Deepfakes: Trick or treat? [J]. Business Horizons, 2020, 63(2): 135–146.

[2] Stamatis K. Artificial Intelligence in digital media: The era of deepfakes [J]. IEEE Transactions on Technology and Society, 2020, 1(3): 138–147.

DOI 10.15302/J-SSCAE-2021.03.004

[3] Tavakoli M, Baldi P. Continuous representation of molecules using graph variational autoencoder [C]. CA: 2020 AAAI Spring Symposium on Combining Artificial Intelligence and Machine Learning with Physical Sciences (AAAI-MLPS 2020), 2020.

[4] Liu H, Zheng X Y, Han J G, et al. Survey on GAN-based face hallucination with its model development [J]. IET Image Processing, 2019, 13(14): 2662–2672.

[5] Brashier N M, Schacter D L. Aging in an era of fake news [J]. Current Directions in Psychological Science, 2020, 29(3): 316– 323.

[6] Jeffcao. The “worries” in the era of Artificial Intelligence: US congressional hearings discuss the risks and countermeasures of “deepfake” [R/OL]. Beijing: Tencent Research Institute, (2019-07- 02) [2020-11-15]. https://www.tisi.org/10852. Chinese.

[7] Tolosana R, Vera-Rodriguez R, Fierrez J, et al. Deepfakes and beyond: A survey of face manipulation and fake detection [J]. Information Fusion, 2020 (64): 131–148.

[8] Zhou X Y, Zafarani R. A survey of fake news: Fundamental theories, detection methods, and opportunities [J]. ACM Computing Surveys, 2020, 53(5): 1–40.

[9] Joseph A D, Nelson B. Adversarial machine learning [M]. Cambridge: Cambridge University Press, 2019.

[10] Mueller S T, Klein G. Explanation in Human-AI systems: A literature meta-review synopsis of key ideas and publications and bibliography for explainable AI [R/OL]. (2019-02-05) [2020-11-15]. Florida: Institute for Human and Machine Cognition Pensacola United States, https://deepai.org/publication/explanation-in-human-ai-systems-a-literature-meta-reviewsynopsis-of-key-ideas-and-publications-and-bibliography-forexplainable-ai.

[11] Karlo D F, Mario B, Nikica H. Explainable artificial intelligence: A survey [C]. Opatija: the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2018), 2018.

[12] Yu G M, Ma S Y. Artificial intelligence improves online public opinion analysis capabilities [J]. Internet Broadcast, 2017 (2): 85–87. Chinese.

[13] Raman R S, Ensafi R. Censored planet: An Internet-wide, longitudinal censorship observatory [R/OL]. Ann Arbor: University of Michigan, (2020-11-10) [2020-11-15]. https://censoredplanet.org/censoredplanet.

[14] Strategic Research on Artificial Intelligence 2.0 in China Team. Strategic research on Artificial Intelligence 2.0 in China [M]. Hangzhou: Zhejiang University Press, 2018. Chinese.