A Thinking of Ethics of AI and Cybersecurity in Sovereignty Perspective

FIT5129 – Xiaohui Ding – 31291252

Exclusive Summary

This report will discuss and analyze the given article written by Timmers (2019) to practice our abilities of critical reading and use of literature research techniques. Meanwhile, it also broadens our knowledge to better understand the current ethical challenges brought by AI-enabled sovereignty in the context of cyberspace and approaches that governments can take.

There are many findings in this report. AI-enabled cyberspace has brought up ethical issues like privacy invasion and residual risks like the cost of lives, thus state sovereignty should ameliorate their cyber laws by considering the AI factor. Strategic partnerships will collaborate on AI technology information sharing but will also divide the world into different camps. That is why we need global common good, which will benefit us all but also hard to achieve. We also find that different countries have different standards of sovereignty costs, so that might be the reason why international conflicts emerge.

In the end, we will infer to our conclusions that how can these three approaches be improved and how to balance sovereignty and other aspects.

Introduction

The purpose of this report is to find a way to improve and integrate risk management, strategic partnerships, and global mutual interests in the context of AI and cybersecurity so that the world can resolve the tense international atmosphere and uphold human rights.

This report will cover the topics of AI, cybersecurity, ethics, and sovereignty. More importantly, it will focus on the relationships between these topics, the potential threats, and policy recommendation.

The Ethical Challenges of AI enabled Risk Management

According to Timmers (2019), in the context of the rapid growth of the internet and modern technology, an AI-based sovereignty level risk management will help cybersecurity experts to analyze malicious cyber behaviours more efficiently by using big-data filtration.

However, along with the application of AI in strategic autonomy, more and more ethical challenges have gradually surfaced. First of all, although AI can be used to monitor massive data, the individuals who are being watched may feel their privacy rights are invaded. For example, people sometimes are enforced to accept this surveillance, otherwise, they might lose their jobs or be blamed for challenging the policy or government (Timmers, 2019). Secondly, citizens may concern about losing control of their data. For example, some people worry about facial recognition will use their personal information without asking for permission. Besides, the lack of transparency will also lead to fear of how decisions are made. An example is a black box inside AI. It is still unclear what logical steps AI takes to get to the results, yet some policymakers still impetuously believe it. It also causes another ethical issue. Shirking all the liability to AI is simple. For instance, if an AI system makes a wrong decision, the people who are originally responsible can easily make an excuse. Last but not least, the human factor will affect the result. For instance, the human-in-the-loop type of machine learning will be assigned a default setting by some people.

Apart from these ethical issues mentioned by Timmers (2019), more challenges should be noticed. Nie and Sun (2016) suggested that using an intelligent big-data model to predict and counter terrorism will strengthen the ability of homeland defence, but this will also bring up many ethical problems. To begin with, the intelligent machine cannot avoid human biases. If the risk score of a person who has a certain race or a certain belief is set to be higher than others by default, then he/she will be more likely to be recognized as a potential terrorist, even if he/she is not. The AI just does what it is told and taught to do, but the innocent people who are accused to be terrorists will be in pain of fear and desperation. As a result, it will damage the personal reputation and deepen the discrimination of race and belief, which is not only an ethical issue but also harms the sovereignty because it will exacerbate the fragmentation of the country. People who misunderstand each other will become irrational and make more hate crime. The wrong accusations made by the AI model will also destroy the credit system of the whole society.

Another thing to worry about is data breaching. According to Solove and Citron (2016), over the past 20 years, data breaches have caused serious legal and financial trouble to many organizations. On a personal level, data breaches can provide opportunities for hackers to easily steal the identities of millions of people. However, nowadays the governments increasingly rely on the utility of AI, and the learning process of AI increasingly relies on the massive amounts of collected personal data. The problems are how to store the data so that it can be well protected from adversaries and who to blame if data leakage happens. When the data is invoked on a national level, citizens will be trapped in an ethical dilemma where they do not know whether it is possible, reasonable, and correct to require governments to treasure it as they do. If a government-controlled database is exploited, usually the individuals will feel a sense of powerlessness. The state governor can of course blame it on technicians, but no one can ensure that things like this will not happen again. Yet we still give our data to governments no matter how unwilling we are.

Residual Risks of AI enabled Cyber Threats

Aside from the ethical issues mentioned above, there are also some residual risks. As mentioned by Bandyopadhyay and Mookerjee (2017), companies normally manage the risks of their cyber assets by evaluating the security of the product itself and purchasing cyber insurance. Thus, it is financially tolerable for risk management to retain some residual risks.

However, non-financial residual risks should also be noticed. Timmers (2019) indicated that on a political or national level, the non-financial residual risks which are caused by AI-decided behaviours might lead to some irreparable consequences. Timmers (2019) gave the example of the lives lost in Wannacry. When the AI detected an urgent cyber threat, the defensive principles embedded in it derived the decision to shut down the electricity to protect the system from virus. Finally, patients who are having surgery were put in an extremely dangerous situation. Timmers (2019) also stated another residual risk which is that the risk management cannot detect all the vulnerabilities behind a system. A backdoor planted in an application or infrastructure is called a kill-switch. Risk management may fail when this kind of backdoor is not found. A kill-switch usually does not take effect immediately, but like a button for launching the missile, it plays a virtual role in international checks and balances. With the blooming of cyber threats, every government has sufficient reasons to carefully invest the use of external programs and try to stay away from residual risks like kill-switch.

Therefore, residual risks in AI-enabled cybersecurity can be defined as the remaining risks of being attacked after the AI-related mitigation operations are performed. From a financial perspective, risk management has a clear range of acceptance of unexpected threats. For instance, Mukhopadhyay et al. (2017) proposed a cyber insurance model to estimate the budget of cybersecurity in risk management. But in the non-financial aspect, risk management usually cannot assess the loss because it is sometimes related to ethics. An example is the recent SolarWinds breaches. It is widely reported that SolarWinds servers were heavily attacked on 13 December 2020, and the most noticeable thing is the leak of weaponry of FireEye (“SolarWinds supply chain breach”, 2021). FireEye possesses many automated cyber warfare weapons, so a breach like this will bring up immeasurable consequences. The adversaries can apply reverse engineering on FireEye's AI-related defensive or aggressive system and utilize the results to optimize their attacks against the AI system. Even though SolarWinds has fixed its vulnerability and FireEye has updated its weaponry, the influence will still be there. The residual risks that are caused by this data breach cannot be quantitively evaluated by money because it put the whole country in danger.

To sum up, residual risks are harmful. Financially, the organizations can choose cyber insurance, but non-financially, the policymakers should consider deeper the costs of lives and credibility.

Three Approaches for Addressing Ethical Challenges in Cybersecurity

Aside from state lead risk management, Timmers (2019) also stated 2 more approaches is often used by governments which are partnerships between countries and universal mutual interest. We will discuss these three approaches in more details.

The first approach is state lead risk management in AI and Cybersecurity. The benefit of using AI-enabled risk management to deal with the increasing international cyber threats is obvious. Governments can monitor their cyber assets, detect a real-time cyber threat, and quickly restore after being attacked (Timmers, 2019). AI can also be used in massive data filtering. For example, assume that a company was attacked. The cyber experts will retrieve the pcap log file of the network traffic and use Wireshark to parse every request. They have to analyze them manually in old days. Now with AI, the specific patterns of hacking are revealed, and AI can take advantage of these patterns to capture malicious behaviours. But there are still some ethical issues and residual risks which we have already talked about in the previous two sections. So, the question is how to deal with the remaining issues and risks. Timmers (2019) recommended that both national cyber strategies and international norms should be considered. Confidence Building Measures (CBMs) are practical ways to achieve these norms. It includes principles like 'do not hurt', 'do not cheat', 'value equality and responsibility'. But CBMs are also quite unrealistic. For instance, United Stated charged Russia for interfering with the 2016 election by cyber espionages which broke the norms of international cybersecurity (Finnemore & Hollis, 2019). Currently, CBMs only succeed between mutually trusted allies for effectiveness. There are still some good measures that states can take to improve risk management, though. The whole process of the AI defensive system should all be well evaluated and controlled. To maintain cyber-resilience, governments should make AI-related cyber behaviours more transparent and standardized. Modifying laws to adapt to cyberspace and correct the fault of the human factor in AI will also be a good policy.

The second approach strategic partnerships are the relationships between mutually trusted governments that will benefit the growth of the economy, science, and the whole society. There are three functionalities of AI in this approach. AI can be used as a part of cyber assets that should be protected, a way of defence in cyberspace, and a powerful weapon in cyber warfare (Timmers, 2019). Strategic partnerships usually exist among the countries that value the same point of view. For example, UKUSA – also called 'Five Eyes' – is a cooperative alliance including intelligent agencies of Canada, the United State, the United Kingdom, Australia, and New Zealand. Shiraz and Aldrich (2019) indicated that this US lead organization has been applying data monitoring for a long time in BRICS countries. This kind of partnership helps to internally share the intelligence more efficiently but will also hurt the trust with other countries or organizations. Thus, there are also some ethical issues in the strategic partnerships approach. Timmers (2019) stated that the EU sometimes set a trade barrier with other countries for protecting privacy. When AI-related cyber business is involved in human rights and trade behaviours, it is sometimes hard to distinguish whether the policy is made for the intention of human rights or pretence of trade barriers. It should be further discussed how to balance the sharing of resources and preventing invasion of privacy in strategic partnerships. At the same time, globalization is another recommended policy for the strategic partnerships approach mentioned by Timmers (2019). Also, all strategic partnerships should take precautions of weaponized AI.

As the third approach, the global common good is very different from the other two. It is defined as a non-governmental centric approach that helps to deal with the issues brought by fast-developing AI and cyberspace (Timmers, 2019). This approach can unite every country in the world to fight against online child pornography or hunt down international criminal organizations. Garrity (2017) believed that the Internet is the next generation of technology as a global public good, which can enable governments to agree on specific ethical consensus so that governments can focus more on their state-centric risk management. For example, fighting sexism, racism, cyberbullying and protecting human rights are the shared responsibility of the world. However, the weakness of the global mutual interest approach is apparent. The competition and conflict between strategic alliances has gradually increased since the arrival of the pandemic. Deglobalism is prevailing. Global common goods currently can only stay on an initiative level, but on the practice level, this is still unrealistic. Therefore, Timmers (2019) suggested that governments should pay more attention to the global common good approach. The private-public intergovernmental cooperation is worthy of being used more widely. In the age of AI, governments should collaborate more rather than constantly slander each other.

In short, the three approaches can be represented by the graph below.

	Definition	Benefit	Issues	Policy Recommend
Risk Management	Identify, protect, detect, defend, recover in AI-enabled cyber threats	Monitoring; big data-based threat detection; real-time response; recovery; helps CERT && big business values	Monitoring: intrusive & coercive; Sense of losing control of their own information; Not transparent; Transfer the blame to system; fallacy of human in the loop.	International norms & values; Practical: whole AI chain; AI-enabled cyber exercise; interoperability, standardization, certification and promoting; legislation
Strategic Partnerships	Working with sufficiently trusted governments in AI-enabled cybersecurity.	Like-minded strategic partners will benefit from collaboration; internal resources sharing	Trade barriers in the guise of human rights; balance of sharing and hiding	Global vision; avoid weaponized AI; update trade policy and cyber law
Global Common Good	The mutual interests that benefit the whole world	Let countries focus on state-centric risk management, fighting crimes and inequality together	Not widely realistically used; not enough international governance	Private-public collaboration; Intergovernmental work; Personal data protection

Sovereignty Costs of Unethical Cyber Attacks

Apart from the approaches mentioned above, Timmers (2019) also put forward a new perspective. Sovereignty costs are trading off or sacrificing individual interests like privacy or even human rights to a certain extent for protecting national legitimacy and state governance. In the age of AI, with the help of cyberspace, state sovereignty is facing more challenges than ever. For example, official powers like to use big data filtering to track terrorist information or apply AI-defensive system on significant infrastructures. But this will also infringe the freedom of speech or hurt lives. A balance between sovereignty and human rights is urgently needed. Intelligent adversaries are challenging state legitimacy, too. State legitimacy can be sorted into internal or external legitimacy. We will use two examples to elaborate on how they are challenged by unethical cyber attacks.

As mentioned by Yang and Liu (2014), China has 538 million users in cyberspace in 2012 yet its cultural influence on mainstream social network service is minimal. It seems that Simplified Chinese users disappeared on YouTube, Twitter, and Instagram. It's all because of the Great Firewall project, which originally got noticed in public eyes On March 30, 2010 because of the fully banning of Google search (Kim &Douai, 2012). The costs of GFW are huge. Since 2010, the Simplified Chinese cyberspace is gradually isolated from the whole world, which has led to many toxic issues. People who are inside the wall easily get irritable and sensitive to those who have different views of the country. The disconnection also affects the communication of culture and technology. Why is China willing to trade off so many things to build the GFW? The reasons are complicated. Due to the structure of its government, China is a sovereignty that needs internal legitimacy more than external legitimacy. The increasing cyber threats including fake news, conspiracy theory and trolling robots may imperil the stability of national security, and China is determined to prevent the split of thoughts at all costs.

Another counterexample is the United States. Ali & Zain-ul-abdin (2020) stated that fake news and conspiracy theories on Facebook had a huge impact on the 2016 U.S. election. This is the curse of lack of supervision. The recommendation algorithm always pushes like-minded things to its users, which sometimes contains unreviewed false information. If this is used by state or non-state intelligent social engineering, the whole election can be controlled. But why the United State is willing to pay the costs? One reason is that the United States is made up of multiple states and multiple races. Normally, people who live in the United States have different believes and views. Excessive censorship of speech can damage the interests of the U.S. political system more than fake news and conspiracy theories. Thus, the United States chooses to trade off its stability.

Summary of Discussions

To sum up, in the age of AI, with the help of cyber, there are plenty of ethical issues and residual risks in state lead risk management. The former includes invasion of privacy, feeling of restlessness about personal information, lack of transparency, transferring responsibility, and the human factor in AI. The latter is defined as the risks after mitigation of cybersecurity and includes the wrong decision made by AI. Thus, Timmers (2019) introduced three approaches to deal with these increasing challenges. State lead risk management which is defined as the control of risks of AI-enabled cybersecurity on a state level will monitor malicious cyber threats by using AI defensive system. Strategic partnerships are defined as the collaboration between like-minded governments in cybersecurity. Some classic examples are EU cooperation, the UKUSA alliance, BRICK partnership. The last approach is the global common good that is defined as the mutual interests in cybersecurity benefiting every country. It consists of fighting for child abuse, international cybercrimes, sexism, racism, and so on. In the end, Timmers (2019) talked about sovereignty costs which are defined as a kind of sacrifice for safeguarding national sovereignty in nowadays AI-related cyberspace.

Conclusion – 5%

In conclusion, with the fast development of AI-related cyberspace, sovereignty is facing many fresh challenges from multiple aspects like ethics, international geopolitics, and state legitimacy. Governments should let go of their prejudices and be committed to realistic confidence building measures rather than shamelessly break the initiatives. It should also be promised not to use AI as a weapon for cyber warfare. Different countries usually have different standards for balancing sovereignty and its costs, which sometimes contribute to an international dispute. Therefore, international neutral organizations have the responsibility to make a global standard stipulating what sovereignty costs can be tolerated and what can not. With the standardization, the international arguments will be more convincing so that the liability will be clearer.

References

Ali, K., & Zain-ul-abdin, K. (2020). Post-truth propaganda: Heuristic processing of political fake news on Facebook during the 2016 U.S. presidential election. *Journal of Applied Communication Research*, 49(1), 109-128. doi:10.1080/00909882.2020.1847311

Bandyopadhyay, T., & Mookerjee, V. (2017). A model to analyze the challenge of using cyber insurance. *Information Systems Frontiers*, 21(2), 301-325. doi:10.1007/s10796-017-9737-3

Finnemore, M., & Hollis, D. B. (2019). Beyond naming and shaming: Accusations and international law in cybersecurity. *SSRN Electronic Journal*. doi:10.2139/ssrn.3347958

Garrity, J. (2017). Getting connected: The internet and its role as a global public good. *Georgetown Journal of International Affairs*, 18(1), 6-8. doi:10.1353/gia.2017.0002

Kim, S. W., & Douai, A. (2012). Google vs. CHINA'S "GREAT FIREWALL": Ethical implications for free speech and sovereignty. *Technology in Society*, 34(2), 174-181. doi:10.1016/j.techsoc.2012.02.002

Mukhopadhyay, A., Chatterjee, S., Bagchi, K. K., Kirs, P. J., Shukla, G. K. (2017). Cyber risk assessment and MITIGATION (CRAM) framework Using Logit AND PROBIT models for Cyber Insurance. *Information Systems Frontiers*, 21(5), 997-1018. doi:10.1007/s10796-017-9808-5

Nie, S., & Sun, D. (2016). Research on counter-terrorism based on big data. *2016 IEEE International Conference on Big Data Analysis (ICBDA)*. doi:10.1109/icbda.2016.7509788

Shiraz, Z., & Aldrich, R. J. (2019). Secrecy, spies and the global South: Intelligence studies beyond THE 'Five eyes' alliance. *International Affairs*, 95(6), 1313-1329. doi:10.1093/ia/iiz200

SolarWinds supply chain breach Threatens government agencies and Enterprises worldwide. (2021). *Network Security*, 2021(1), 1-3. doi:10.1016/s1353-4858(21)00001-5

Solove, D. J., & Citron, D. K. (2016). Risk and anxiety: A theory of data breach harms. *SSRN Electronic* Journal. doi:10.2139/ssrn.2885638

Timmers, P. (2019). Ethics of AI and Cybersecurity when sovereignty is at stake. *Minds and Machines*, 29(4), 635-645. doi:10.1007/s11023-019-09508-4

Yang, Q., & Liu, Y. (2014). What's on the other side of the great FIREWALL? Chinese WEB users' motivations for bypassing the internet censorship. *Computers in Human Behavior*, 37, 249-257. doi:10.1016/j.chb.2014.04.054