June, 2021

JDLA

Study Group 'AI governance and its Evaluation'

Report on the Session #12

1. Introduction

The Japan Deep Learning Association establishes study groups as a forum for deepening knowledge and discussing domestic and international policy trends related to artificial intelligence (hereafter AI) and Deep Learning (hereafter DL). This study group, "AI Governance and its Evaluation," defines "governance" as a system of management and evaluation by various actors, and launched a study group in July 2020 to investigate what forms of governance are possible and conduct a year-long study to help build trustworthy AI systems.

In the 12th session (April 22, 2021), Prof. Jun Sakuma of University of Tsukuba and Mr. Yusuke Nirahara of BrainPad Inc. presented topics under the theme of "Security towards AI Governance".

This report is a reconstruction of the topical presentations and the discussions of the study group participants.

2. AI security, privacy and trust issues

Prof. Sakuma gave a presentation on "AI Security, Privacy and Trust issues".

Conditions for trustworthy AI Decision Making

In recent years, with the development of deep learning technology, the accuracy of image recognition by AI has been increasing, and when focusing on simple purposes such as numeric classification, the accuracy of AI recognition has surpassed that of humans. AI is now being applied to more complex decision-making tasks such as medical diagnosis and automated driving.

However, when AI is used as an expert, as in the case of medical diagnostic AI, there is a high risk that an erroneous decision by AI will result in a serious decision that could affect the life of a human being. Therefore, the results of AI decision making need to be not only correct, but also ethically sound based on the explanation and support of the rationale and process.

This is because even in the case of humans, people who need to make decisions that have a significant impact on human lives, such as doctors and lawyers, are expected to have high professional ethics and are required to adhere to ethics through laws, professional guidelines, and internal constraints. Therefore, even when AI takes on jobs

that require high professional ethics, we need to make sure that the results of AI decisions are ethically sound.

Prof. Sakuma cites the following Table 1 as conditions for trustworthy AI decision making.

Table 1: Conditions for trustworthy decision makings by AI[1]

Minimum required Conditions	Expert-level quality	
	Compliance with Laws and Regulations	
Constraints	Resilience to environmental changes and attacks	
	Respect for human rights	Respects for human autonomy
		Privacy-preserving
		Fairness-awareness
Verification	Explainability of the basis for the decision	
	Traceability of the decision-making process	

- ✓ Constraints
  It is a condition that should be considered for quality, even if it somewhat limits the goal of profit maximization.
- ✓ Resilience to environmental changes and attacks
  See "Attacks on AI" below.
- ✓ Respects for human autonomy
  Respect for the human autonomy must be a condition for AI to make decisions that affect human lives, such as in medical treatment or court cases, because depriving humans of their right to self-determination may cause serious problems.
- ✓ Privacy-preserving
  In the development and use of AI, it is necessary to consider the technical aspects of AI, such as the fact that data collected without the full consent of individuals/ companies should not be used, and that personal information should not be exploited from AI models.
- ✓ Fairness-awareness
  See "Fairness-awareness" below.
- ✓ Explainability of the basis for the decision
  See "AI Security," below.
- ✓ Traceability of the decision-making process
  When an AI makes a wrong decision, it is essential that the decision-making process can be tracked by humans to see how the wrong decision was made.

[1] Excerpts from the public materials of this study group.

If AI services are designed without considering the conditions for trusted decision makings by AI, there is a risk of making third parties unhappy, which is not expected by AI service providers. It is vital to keep the aforementioned perspectives in mind when developing AI services.

Attacks on AI

There is an attack on AI called adversarial examples, which add subtle noises to the input data, causing the AI model to make false predictions. For example, when noise is cleverly added to an image of a panda, the AI recognizes it as a gibbon instead of a panda. It turns out that adversarial examples can be created for any image.

Adversarial examples can also be created for audio, which contains subtle noise to make the AI recognize it as something else. For example, when a sound that sounds music to humans contains a minute noise, the voice recognition AI recognizes it as a voice command calling for a fire engine. Audio adversarial examples can be broadcast from speakers, radios, etc., and thus are highly diffusible and can be used to reach and forcibly manipulate the target device.

The adversarial examples also attack explainability, feeding the AI with image that contain a faint noise to make it give a completely different explanation. If the AI behaves unstable as described above, it may not be trusted by AI service users even if the performance of the AI itself is high. Therefore, stability against environmental changes and attacks is necessary for trustworthy AI decision makings.

Fairness-awareness

AI decision making would impact human lives (e.g., hiring, credit scoring, admissions, insurance rating). Some AI have already been making some of the decisions that humans make. Therefore, fairness needs to be considered in AI decision making.

The reason AI makes decisions that cause discrimination is due to its dependence on human attributes. For example, if hiring AI depends on gender, it may decide to hire males and reject females, even if their performance and background data are similar. This is referred to as gender discrimination by AI. However, regardless of whether the decision-making entity is human or AI, it is necessary to define discrimination and consider fairness.

COMPAS, an algorithm that predicts a defendant's likelihood of recidivism, has been adopted by several states in the U.S. to assist judges in making decisions on bail amounts, sentencing, and other matters. In 2015, however, Julia Angwin pointed out that COMPAS's predictive results were subject to racial bias. She found out that in some

cases, COMPAS assessed lower recidivism risk scores for whites than was actually the case, and higher scores for people of color. Although COMPAS itself is a simple algorithm that cannot be called AI, it suggests that there is an issue of how to evaluate the fact that discrimination is caused by statistical judgment and how to ensure fairness. One of the measures of fairness is demographic parity, is the conditional distribution of decision makings with respect to the consideration attribute (e.g., gender) and the gap with respect to the consideration attribute value (e.g., male/female). One measure to ensure fairness is to make this gap smaller and let it learn to achieve the fairness indicator.

AI Security

Communications security is security for the simple task of communicating information, but no perfect defense against unauthorized access to communications has yet to be found. For many years, there has been a tit-for-tat situation in which defenders take countermeasures against attacks, and attackers find vulnerabilities in those countermeasures and attack them. As a result, continuous research on security measures has become the essence of maintaining a high level of security.

Security measures for AI are even more complex than communications security measures; the key point in AI security is that attacks against AI are often carried out by AI, so defensive measures must also be made by AI.

Note that just as in the case of communications security, AI security will continue to be a battlefield as well.

3. Cyber threats related to AI and its countermeasures

Next, Mr. Nirahara spoke on the topic of "Cyber Threats Related to AI and its Countermeasures".

Security for AI systems

There are six points in an AI system that can be targeted for attack: input data in the learning process, algorithms during information processing, learned models, original data in the inference process, models, and output results.

An example of an attack on the original data in the inference process is to change the original image by one pixel, causing the AI to make a wrong decision in image recognition. As another example of an attack on output results, although not AI, is the cyberattack on the supervisory control and data acquisition (SCADA) system of Iran's nuclear facilities using Stuxnet. In this attack, the rotation cycle of the centrifuges was falsified, and the centrifuges were destroyed over a period of several months to six

months by making the monitoring monitor show that the centrifuges were normal. This cyberattack on the nuclear facility was a so-called zero-day attack. The vulnerability can be eliminated by applying a corrective patch, but zero-day attacks target unknown vulnerabilities, making it difficult to prevent attacks beforehand. The case suggests that when considering the security of AI systems, it is necessary to consider the security of the entire software that incorporates the AI, not just the AI model.

Cyber Resilient Software

There is an initiative to solve the issue of increased redelivery by determining the home status of residents based on the amount of electricity used by smart meters. On the other hand, digitalization of home status may pose a risk of kidnapping, assassination of important persons, theft, burglary, etc. Since there have been incidents of theft by SECOM security guards,[2] full-scale commercialization of the service is required after taking into account the possibility of incidents such as the aforementioned occurring through the use of smart meters. It is also necessary to consider measures to deal with the risk of data leakage in the supply chain related to the service provision by telecommunication companies, power companies, delivery companies, analysis companies, etc. While users can enjoy the benefit of reduced redeliveries by using smart meter data, they must be aware that they themselves are exposed to the aforementioned risks, and it will be essential to operate services based on consensus.

However, in reality, it is sometimes difficult to take full-scale countermeasures against cyberattacks until after the actual damage has been occurred. For example, Google has been working on building a Zero Trust Network[3] since January 2010, when the company announced that its intellectual property had been stolen due to hacking from China.[4] The Zero Trust Network is based on the premise that all systems must be hacked and all communications must be authenticated (Zero Trust), instead of the "cyber security" concept of building a wall to protect and defend against cyberattacks. This should be done in conjunction with the concept of "cyber resilience," which focuses on recovering from cyber-attack damage as soon as possible and gaining stronger super-resilience. In recent years, the U.S. Department of Defense and the Davos Forum have been focusing on "cyber resilience" as a matter of concern.

DevSecOps[5] is useful for developing cyber resilient software. It applies security at all

---

[2] https://www.asahi.com/articles/ASMDB5CTPMDBPIHB02C.html (in Japanese)

[3] A network security concept. The approach is to inspect, log, and examine the traffic of all devices, assuming that they should not all be trusted by default.

[4] https://googleblog.blogspot.com/2010/01/new-approach-to-china.html

[5] DevOps is a set of practices based on the concept of software development (Dev) teams and IT

phases of the software lifecycle: planing, service design, development, testing, and operations. Specifically, in the planning phase, it considers what kind of attack threats exist; in the service design phase, it detects the possibility of AI service users being compromised; in the development phase, it secures coding and analyze source code vulnerabilities; in the testing phase, it conducts automated vulnerability testing and countermeasures against vulnerabilities identified by attacks from white hackers; in the operation phase, regular security scans and security audits are conducted, and open source is updated if vulnerabilities are found in the open source.

Recently, the concept of security-conscious AI development called "MLSecOpes," which combines DevSecOps and MLOps[6], has been advocated mainly in North America. In addition, the Adversarial ML Threat Matrix, an open source framework for detecting, responding to, and remediating threats against ML (machine learning) systems, has been released by Microsoft, IBM, Carnegie Mellon University, and other volunteer communities. [7]

Hybrid Threats

The major threats to international security in the future will be hybrid threats that combine methods of attack on physical space (land, sea, and air), information space (cyber and cognitive space), and outer space.

Hybrid threats came to prominence in the wake of Russia's operations conducted during its annexation of Crimea in 2014, and warfare across domains has come to be referred to as hybrid warfare. Hybrid warfare is one of the theories of military strategy which combines conventional warfare [8], irregular warfare [9] and cyberwarfare with other influencing methods. The annexation of Crimea by Russia was the result of a referendum held by the Crimean Parliament in which more than 95% of the votes cast in favor of the annexation of the autonomous Republic of Crimea from Ukraine and annexed it to the Russian Federation.[10]

＜Hybrid threats seen during annexation of Crimea＞

- ✓ Blocking communication lines between Ukraine and Crimea

operations (Ops) teams complementing each other's missions to increase the usefulness of systems to the business and to ensure rapid and reliable implementation and operation. DevSecOps is the combination of DevOps development and operations with the addition of security.

[6] Based on the DevOps approach, it refers to a method that integrates the development and operation of machine learning (ML) models and manages a series of lifecycles from development to operation.

[7] https://github.com/mitre/advmlthreatmatrix

[8] A form of warfare conducted by a trained military force organized by the state.

[9] A form of conflict that cannot be categorized as conventional warfare, such as a conflict in which a people take up arms voluntarily or with the help of a foreign power.

[10] The annexation of Crimea is not considered to have received international approval.

- ✓ Using social networking to spread information about demonstrations and riots in Ukraine and information that may cause unrest among Russian residents in Crimea.
- ✓ DDoS attacks on websites
- ✓ Occupation of airports, government buildings, and the Crimean Provincial Council by vigilantes known as the Little Green Men

Focusing on space, positioning satellites (e.g., GPS), communication satellites, meteorological satellites, and early warning satellites communicate between space and the earth, thus space is also a target of cyber security. Therefore, it is necessary to consider the security risks (attacks on satellites[11]) of the internal mechanism of satellites, earth-to-satellite and satellite-to-satellite. Furthermore, given that autonomous control functions based on AI technology are essential for spacecraft and rockets in space development, such as the asteroid explorer Hayabusa, it is necessary to take measures against these security risks in future space development.

The idea of security resilience under hybrid threats requires a multi-domain (land, sea, air, cyberspace, cognitive world, space, AI is also part of it) security resilience mindset; AI systems are important, but they are only part of the total software. The response to hybrid threats needs to be considered based on the national security policy.

Cyber Grand Challenge

In August 2016, DARPA[12] hosted the Cyber Grand Challenge (CGC), the world's first competition to create automatic defense systems capable of identifying software flaws, formulating patches and deploying them on a network, and attacking their opponents' weakness.

At the 2016 CGC, a computer system "Mayhem" created by a spin-off team of the Carnegie Mellon University known as ForAllSecure won the competition. Mayhem uses an automated software testing technique called fuzzing to automatically test for program vulnerabilities. It involves automatically inputting amounts of data to the test subject identify vulnerable areas.

＜Types of Fuzzing＞

- ✓ Random fuzzing: Feeds random inputs into a program

[11] The types of attacks on satellites include (1) arm, scientific spray, net, high-power microwave, and jamming attacks from military satellites, (2) laser irradiation of optical satellites from the ground, (3) attacks by anti-satellite attack missiles, (4) spoofing and jamming attacks on ground inter-satellite communications, and (5) cyberattacks on ground control facilities.

[12] Abbreviation for Defense Advanced Research Projects Agency

- ✓ Template fuzzing: Feeds inputs based on a set of manually prepared templates into a program.
- ✓ Generational fuzzing: Feeds auto-generated inputs into a program.
- ✓ Guided fuzzing: Automatically generates inputs and feeds the next inputs into a program while learning from the results. The effectiveness of each test case is scored and reinforcement learning is performed.

Fuzzing is a technique that is routinely used by Google and Microsoft for vulnerability detection. Mayhem, which combines fuzzing and other testing techniques, has since been adopted by the U.S. Department of Defense and is being used as a highly accurate vulnerability detection AI system. How to develop such AI for cyber defense is also becoming a critical point in the future security of the cyber world.

4. Discussion points in the question and answer session

In the 12th session, the contents of the "Security towards AI governance" were discussed. The following questions and answers were raised based on the topics discussed.

Security concepts for AI systems

- ✓ It is necessary to be aware that the risk assessment results of an AI model alone may be different from those of the entire AI system.
- ✓ Since AI is a part of a system, it is necessary to set up system security by utilizing general system security frameworks and standard procedures. However, since it is difficult to address all security risks of AI, it is necessary to build AI systems with the knowledge to assume about potential threats and to make assumptions to be able to respond to threats when they occur.
- ✓ It is necessary to recognize the possibility of AI services being attacked and to assume the purpose and target of AI services being attacked. In addition, it should be recognized that the purpose of cyberattacks is not only for amusing crimes, but also for military purposes, monetary purposes, etc.
- ✓ Security measures need to be changed depending on the scenario of an attack on AI services; what are the issues that pose the greatest risk to AI services and the likelihood of those issues occurring need to be considered.
- ✓ Security measures are often considered by assuming the worst case. The approach of assuming and taking measures based on the usage pattern does not necessarily achieve perfect security. However, since it is not realistic to take measures against all cyber security risks, it is possible as a security strategy to take measures starting

from the areas where vulnerabilities are likely to affect the external environment in light of the service usage patterns.

- ✓ The level of cyber security risk measures for their services is something that even GAFAM[13] personnel are struggling with. Prioritization of security measures and compromises that allow for some degree of risk are also necessary for business operations.
- ✓ If the behavior of a system that deviates from expectations is a bug, then in the case of AI, any behavior that deviates from expected decision-making, including issues of robustness and fairness, can be considered a bug, so issues caused by AI mechanisms can also be considered an issue within the scope of "security."

Education on AI Security

- ✓ It is important to provide education to increase the number of researchers on AI security issues in universities. Currently, there are not many universities that have departments in the computer science area.
- ✓ With the development of information technology, the field of security as an academic discipline is also developing, and as ML/DL develops further, it is expected that a certain percentage of researchers in conventional laboratories on security will be engaged in research on AI security (research on ML/DL security).
- ✓ It is necessary to conduct research on AI security not only for the security of AI alone, but also in conjunction with cyber and military strategies.

Attacks on AI

- ✓ We should make the assumption that AI is not perfect, because even if AI is not attacked, AI misjudgments will not be zero, and the results of AI decisions will not be 100% correct.
- ✓ In audio adversarial examples, it is possible to deliberately develop noises such as “call the fire department”.
- ✓ Audio adversarial examples can also attack using the ultrasonic bands that are inaudible to humans.
- ✓ Students with specialized education are capable of creating adversarial examples. Although a certain level of ML/DL knowledge is required, the creation of adversarial samples itself is not difficult. However, it is difficult to create the conditions for spreading the adversarial examples to the external environment, and in the case of audio adversarial examples, it is not immediately possible to spread the adversarial

[13] GAFAM is an acronym for the five big American technology companies: Google, Apple, Facebook, Amazon and Microsoft.

examples to other people's smartphones. Adversarial examples are created on the assumption that they will access the speech recognition model, but the specification does not allow for easy access to the speech recognition model of smartphones.

✓ Attacks on AI services are not only technical attacks on AI models, but also human attacks, so even if we try to respond technically, we may be threatened artificially.

How to Increase Security Resilience

✓ It is essential to recognize the fact that security risks will always exist. For example, if you use Google Workspace or Microsoft Office365, you can see that you are exposed to security risks from events such as spam emails being automatically excluded, or device authentication being used to reject unauthorized logins to the system.

✓ Some people do not upgrade their operating systems for fear of the impact such as the possibility that the system may stop working, but it is important to apply patches that resolve known vulnerabilities, and it is also important to establish an operational system to do so.

✓ It is necessary to visualize security risks and accumulate organizational measures and decisions to counter those risks.

Positioning security in businesses

✓ In fact, in digital transformation, profit generation is a priority and security is an afterthought.

✓ It is a business decision to prioritize between improving system functions and addressing security vulnerabilities, and it is required to position security as part of CRM (customer relationship management) and service branding and make decisions on a business level.

We will continue to discuss AI governance in Japan and abroad through this study group.

Written by Yuki Kiyomi
Translated by Michiko Shimizu

<Outline of the 12th Session of the Study Group>

Date & Time: Thursday, April 22, 2021, 14:30-16:30 (Zoom)

Agenda:

- Topical presentations:
  - "AI Security, Privacy and Trust issues" provided by Prof. Jun Sakuma (University of Tsukuba)
  - "Cyber Threats Related to AI and its Countermeasures" provided by Mr. Yusuke Nirahara (BrainPad Inc.)
- Question and answer session / discussion