ACICE
ADMM Cybersecurity and Information Centre of Excellence

UPDATE ON
THE CYBER DOMAIN

Issue 11/23 (November)

Rise of Artificial Intelligence – Extinction of Mankind?

OVERVIEW

1. Experts have warned that Artificial Intelligence (AI) could lead to the extinction of humanity. The statement published on the website of the Centre of AI Safety reads "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." The risks of AI are real and there is endless potential for their misuse if they are not managed properly.

"To harm humans, AIs wouldn't need to be any more genocidal than someone removing an ant colony on their front lawn. If AIs are able to control the environment more effectively than we can, they could treat us with the same disregard." **– According to the Centre of AI Safety**

RISKS TO HUMANITY

2. AI is now poised to become a powerful technology with destructive potential, some described as almost similar to nuclear weapons, which could bring about catastrophic events with devastating consequences for vast numbers of people. Some sources of catastrophic AI risks are outlined below.

3. **Malicious Use:** Threat actors could intentionally harness powerful AI to cause large spread harm. Specific risks include bioterrorism enabled by AIs that can help humans plan an attack with a biological weapon. For example, the attempt by the Japanese Aum Shinrikyo cult in the 1990s to use botulinum toxin as a bio-weapon to kill Japanese leaders in the Japanese parliament failed because of a lack of understanding of the toxin. However, AI can quickly bridge the knowledge gaps and identify any potential weaknesses to the attack. Other examples include the use of AI tools such as WormGPT and FraudGPT by threat actors for nefarious purposes such as phishing campaigns.

4. **AI Arms Race:** In the 1950s and 1960s, the fear that the Soviet Union was leading in the ballistic missile development spurred the United States to scale up their missile technology and triggering a nuclear arms race. Geo-political tensions and competition may force militaries to use AI to develop autonomous weapons. This could enable a new kind of

automated warfare where there is a possibility of losing control of the AI systems, leading to an existential risk such as human extinction or irreversible global catastrophe.

5. **AI Corporate Race:** Competition could pressure nations and corporations to rush the development of AIs and cede control to AI systems. Corporations will face similar incentives to automate human labour and prioritise profits over safety, potentially leading to mass unemployment and dependence on AI systems. According to a report by Goldman Sachs, AI could replace the equivalent of 300 million full time jobs with a "significant disruption" on the horizon for the labour market.

6. **Organisational risks:** The Chernobyl nuclear disaster in 1986 was due to an accumulation of risk factors such as a flawed reactor design, inadequately trained personnel, non-compliance with operational procedures and lapse in safety protocol. Similarly, catastrophic accidents can happen if organisations that develop and deploy advanced AIs do not have a strong safety culture.

7. **Rogue AIs:** Rogue AI is often used to describe an artificial intelligence that commits potentially dangerous acts and can bring with it varying levels of severity, threats or harm. A common and serious concern is that AI may become sufficiently autonomous to set its own goals which do not align anymore with humanity's well-being (or their creator's will).

***"AI could lead to civilisation destruction"* – Elon Musk, Tesla CEO**

MANGAGING AI RISKS

8. AI is growing at a rapid pace which threatens to grow out of control if left unchecked. Hence, there is a pressing need to develop some form of international standards on the use of AI and to mitigate the potential risks. This could encompass technical aspects as well as the ethical and policy dimensions of responsible AI.

***"I would ask these 350 people and the makers of AI — while we're trying to put a regulatory framework in place — think about self-regulation, think about what you can do to slow this down so we don't cause an extinction event for humanity. If you actually think that these capabilities can lead to extinction of humanity, well, let's come together and do something about it."* – According to Jen Easterly, director of the Cybersecurity and Infrastructure Security Agency (CISA)**

9. **Access Restrictions:** AIs might have dangerous capabilities that could do significant damage if used by malicious actors. To mitigate this risk, it may be necessary for organisations to restrict user access to dangerous system capabilities by only allowing controlled interactions with those systems through cloud services and conducting user screenings before providing access.

10. **Legal Liability:** General-purpose AIs can be fine-tuned and prompted for a wide variety of downstream tasks, some of which may be harmful and cause substantial damage.

To reduce risk of that happening, AI companies should be liable for harms that they could have averted through more careful development, testing, or standards.

11. **International Cooperation:** To prevent escalatory AI race, international coordination and cooperation is needed to set standards and agreements as well as build trust on responsible AI development. Cooperation could be accomplished via informal agreements, international standards, or international treaties regarding the development, use, and monitoring of AI technologies as well as the ethical use of AI.

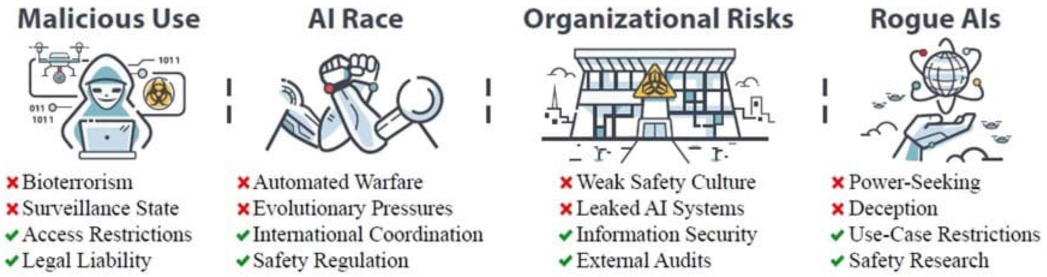
12. **Safety Regulations:** Having safety regulations will provide AI developers, deployers and users with clear requirements and obligations regarding specific uses of AI. These regulations will guarantee the safety and fundamental rights of people and businesses when it comes to AI.

13. **Information Security:** Cybersecurity is critical in ensuring that information on AI is safeguarded. Threat actors will attempt to steal data and infiltrate organisations for nefarious reasons. Information security measures in proportion to the value and risk level of their intellectual property (IP) are needed to keep them secured.

14. **External Audits:** Organisations could consider engaging external audits to ensure the effective, fair and transparent development of AI systems throughout their lifecycle. External audits can also check for issues such as bias assessments, safety, privacy and transparency risks before the deployment of the product.

15. **Use-Case Restrictions:** Certain use cases of AI are riskier than others and organisations should not deploy AIs in high-risk settings until safety has been conclusively demonstrated. AI systems should not accept requests to autonomously pursue open-ended goals requiring significant real-world interaction. For example, it would be disastrous should AI start removing humans if it were asked to "reduce human-made noise as much as possible" since humans are the source of the noise. AI systems should also not be deployed in settings that would make shutting them down extremely costly or infeasible, for example in critical infrastructure such as power stations or dams.

16. **Safety Research:** Reliable mechanisms should be in place to ensure that AI systems do not act deceptively. This would require a strong understanding of AI system internals, sufficient to have knowledge of a system's tendencies and goals; these tools require safety research.



Categories of AI risks and mitigating measures. Source: (Center for AI Safety, 2023).

EFFORTS TOWARDS SAFEGUARDING HUMANITY’S FUTURE

17. Governments and AI developers have recognised the potential risks that AI may bring and are working towards a common approach to identifying them and to mitigate them. The international community has banded together to cooperate on AI to promote inclusive economic growth, protect human rights and to foster public trust and confidence in AI systems to fully realise their potential. The Bletchley Declaration, agreed at the United Kingdom’s AI Safety Summit on 1 Nov 2023, is a good example of international cooperation towards addressing AI governance and risks. The Declaration recognises the need for AI development and for its use to be human-centric, trustworthy and responsible. The European Union is in the process of regulating the AI Act, which will ensure better conditions for the development and use of AI.

CONCLUSION

18. AI has the potential to transform the way we work, play and live and the race is ongoing to see who will make the next breakthrough. A common approach on how to manage this race will ensure that AI can make the world a better place, not worse off.

***“We’re now in the age of AI. It’s analogous to uncertain times before speed limits and seat belts.”* – According to Bill Gates, Former CEO of Microsoft**

Contact Details

All reports can be retrieved from our website at www.acice-asean.org/resource/.

For any queries and/or clarifications, please contact ACICE, at ACICE@defence.gov.sg.

Prepared by:
ADMM Cybersecurity and Information Centre of Excellence

••••

REFERENCES

News Articles

1. Statement on AI Risk | CAIS
[Link: https://www.safe.ai/statement-on-ai-ris]

2. AI Risks that Could Lead to Catastrophe
[Link: https://www.safe.ai/ai-risk]

3. A.I. automation could impact 300 million jobs – here's which ones
[Link: https://www.cnbc.com/2023/03/28/ai-automation-could-impact-300-million-jobs-heres-which-ones.html]

4. AI chatbots could help plan bioweapon attacks, report finds
[Link: https://www.theguardian.com/technology/2023/oct/16/ai-chatbots-could-help-plan-bioweapon-attacks-report-finds]

5. Why Do We Need AI Auditing and Assurance?
[Link: https://www.holisticai.com/blog/why-do-we-need-ai-auditing-and-assurance]

6. A Race to Extinction: How Great Power Competition Is Making Artificial Intelligence Existentially Dangerous
[Link: https://hir.harvard.edu/a-race-to-extinction-how-great-power-competition-is-making-artificial-intelligence-existentially-dangerous/]

7. Chernobyl | Chernobyl Accident | Chernobyl Disaster - World Nuclear Association
[Link: https://world-nuclear.org/information-library/safety-and-security/safety-of-plants/chernobyl-accident.aspx#]

8. The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023
[Link: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023]

9. EU AI Act: first regulation on artificial intelligence
[Link:https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence]