World Academy of Science, Engineering and Technology
International Journal of Law and Political Sciences
Vol:18, No:10, 2024

The Challenge of Assessing Social AI Threats

Authors : Kitty Kioskli, Theofanis Fotis, Nineta Polemi

Abstract : The European Union (EU) directive Artificial Intelligence (AI) Act in Article 9 requires that risk management of AI systems includes both technical and human oversight, while according to NIST_AI_RFM (Appendix C) and ENISA AI Framework recommendations, claim that further research is needed to understand the current limitations of social threats and human-AI interaction. AI threats within social contexts significantly affect the security and trustworthiness of the AI systems; they are interrelated and trigger technical threats as well. For example, lack of explainability (e.g. the complexity of models can be challenging for stakeholders to grasp) leads to misunderstandings, biases, and erroneous decisions. Which in turn impact the privacy, security, accountability of the AI systems. Based on the NIST four fundamental criteria for explainability it can also classify the explainability threats into four (4) sub-categories: a) Lack of supporting evidence: AI systems must provide supporting evidence or reasons for all their outputs. b) Lack of Understandability: Explanations offered by systems should be comprehensible to individual users. c) Lack of Accuracy: The provided explanation should accurately represent the system's process of generating outputs. d) Out of scope: The system should only function within its designated conditions or when it possesses sufficient confidence in its outputs. Biases may also stem from historical data reflecting undesired behaviors. When present in the data, biases can permeate the models trained on them, thereby influencing the security and trustworthiness of the of AI systems. Social related AI threats are recognized by various initiatives (e.g., EU Ethics Guidelines for Trustworthy AI), standards (e.g. ISO/IEC TR 24368:2022 on AI ethical concerns, ISO/IEC AWI 42105 on guidance for human oversight of AI systems) and EU legislation (e.g. the General Data Protection Regulation 2016/679, the NIS 2 Directive 2022/2555, the Directive on the Resilience of Critical Entities 2022/2557, the EU AI Act, the Cyber Resilience Act). Measuring social threats, estimating the risks to AI systems associated to these threats and mitigating them is a research challenge. In this paper it will present the efforts of two European Commission Projects (FAITH and THEMIS) from the HorizonEurope programme that analyse the social threats by building cyber-social exercises in order to study human behaviour, traits, cognitive ability, personality, attitudes, interests, and other socio-technical profile characteristics. The research in these projects also include the development of measurements and scales (psychometrics) for human-related vulnerabilities that can be used in estimating more realistically the vulnerability severity, enhancing the CVSS4.0 measurement.

Keywords : social threats, artificial Intelligence, mitigation, social experiment
Conference Title : ICCCC 2024 : International Conference on Cyberlaw, Cybersecurity and Cybercrime
Conference Location : Osaka, Japan
Conference Dates : October 28-29, 2024