CYBERSECPRO

Human Factors' benchmarking in ensuring trustworthiness of AI systems: Challenges and Research Paths

Nineta Polemi
Cybersecurity Research Lab
University of Piraeus

Kitty Kioskli
Trustilio BV

ENISA
AI CYBERSECURITY CONFERENCE
7 June, Brussels
enisa
EUROPEAN UNION AGENCY FOR CYBERSECURITY

(based on the work published in the 14th International Conference on Applied Human Factors and Ergonomics -AHFE 2023)

Trustworthy AI

AI trustworthiness is considered as the confidence that AI systems behave within specified norms, as a function of the following characteristics:

a) **Technical** (e.g. accuracy, robustness, reliability)

b) **Socio-technical** (e.g. explainability, managing bias, transparency, security, privacy), and

c) **Guiding Principles** (e.g. accountability, reliability, environmental well-being, diversity, fairness, traceability).

Any **AI system is a socio-technical system** with technical, socio-technical, and guiding principle characteristics

AI threats

CYBERSECPRO

- Threats of AI-systems are events/causes/incidents that negatively impact the trustworthiness and their characteristics, thus:

AI threats can be classified as:

➢technical threats (e.g. loss of accuracy);

➢socio-technical threats (e.g. loss of explainability),

➢loss of guiding principles (e.g. loss of accountability)

Measuring AI non-technical threats/ vulnerabilities/risks

CyberSecPro

- Socio-technical and guiding principles threats and vulnerabilities cannot be uniformly identified or measured since not all people have the same level of understanding or learning or behaving or perceptive of notions like bias, fairness, equality etc.

- AI threat assessment requires is the estimation of all type of threats. Hoverer **we do not have scales for non-technical threats**

- The assessment, design and implementation of AI-systems will rely upon the understanding and **modelling humans' profiles (anonymously)**   that include learning, behavioural, psychological, cognitive characteristics, ethical values and patterns in an anonymous manner (to protect their privacy).

- Profiling is also important in understanding the adversaries in the AI- operational environment in order to compute the severity of **socio-technical vulnerabilities**, estimate the **attack potential** that will improve our mitigation strategies.

SOCIO-TECHNICAL CYBERSECURITY SCALES AND MEASUREMENTS

- Limitations of current security vulnerability measurement systems (e.g. CVSS).
- These systems do not consider human factors, such as the psychological and behavioral characteristics of attackers and defenders.
- More realistic estimation of the vulnerability of the system is achieved if we can forecast the attackers' profiles.
- Further research is needed to develop accurate measurements and to evaluate the system's accuracy and objectivity in cyber operations.
- The interdisciplinary research involving cyber psychologists, behavioral analysts, and cyber professionals could advance the CVSS3.1 system.

CYBERSECURITY OPERATORS PERCEPTION ON AI-SOCIAL THREATS

- Risk assessment or Incident handling practices rely on the operator's profiles and their understanding of concepts like bias, fairness, equality, and ethics.
- The operators’ profiles and values impact the socio-threat measurements and handling procedures.
- Further research should focus on behavior-change interventions using co-design approaches, examining factors that influence HAI-cybersecurity teams.

BEHAVIOUR CHANGE IN AI-BASED CYBERSECURITY OPERATIONS

- Research on Human AI Interaction (HAI) in cybersecurity operations needs to be further studied.
- The effectiveness of decision-making tasks during incidents with AI assistance needs to be evaluated.
- The operators' trust and confidence in teammates and AI-assistance play an important role in the effectiveness of cybersecurity practices.
- Behavioural change processes to improve the effectiveness and acceptance of HAI interaction are often neglected.
- Future research should evaluate methodologies to assess the efficiency of HAI teams, develop measurements and design targeted, innovative interventions to improve cognitive factors during cybersecurity practices.

CONCLUSIONS

- Building bridges between cyber engineers, cyberpsychology researchers, behavioral and social scientists is essential for effective cybersecurity AI practices.
- The study of human factors is necessary for effective cybersecurity practices and operations, from measuring profiles and risks to managing security incidents to embracing security policies of AI systems and train AI operators and practitioners.
- Model attackers profiles and build psychometric questionnaires and measurements
- Training on human factors that impact the AI cybersecurity

CYBERSECPRO

Kitty Kioskli, PhD

trustilio B.V.

kitty.kioskli@trustilio.com

trustilio.com

Professor Nineta Polemi

University of Piraeus,
Dpt. of Informatics, Cybersecurity Lab

dpolemi@gmail.com