British Embassy Kuwait

المركز الوطني للأمن السيبراني
National Cyber Security Center

cerc 2024

THE 6TH CYBERSECURITY EDUCATION & RESEARCH CONFERENCE

The Need for Ethics in AI-driven Behavioural Interventions – Tackling Social Engineering and Human Error Ethically

Dr Konstantinos Mersinas

Associate Professor

Information Security Group, Royal Holloway, University of London

konstantinos.mersinas@rhul.ac.uk

KUWAIT COLLEGE OF SCIENCE & TECHNOLOGY
كلية الكويت للعلوم والتكنولوجيا

جامعة الكويت
KUWAIT UNIVERSITY
•1966•

UK Science & Innovation Network

125 عاماً من الشراكة الكويتية البريطانية
125 YEARS OF KUWAITI-BRITISH PARTNERSHIP

People

Mr Divine Chana Chupkemi - Royal Holloway, University of London

Dr Maria Bada - Queen Mary, University of London

Prof Steven Furnell – University of Nottingham

Dr Konstantinos Mersinas – Royal Holloway, University of London

Structure

Behaviour and behaviour change

AI solution for human risk

Cybersecurity behaviour change ethical principles

* Ponemon Institute (2019)

The human factor

- Attribution of security breaches to human factors: 24%*
- Significant component of the overall attack surface
  - Social engineering
  - Human error
- AI: need for human in the loop

* Ponemon Institute (2019)

Behavioural interventions

- Individuals behave in non-desirable (non-secure) ways
  → behaviour change is needed
- Cybersecurity behaviour change (CBC) refers to:
  'any modification in the behaviors of individuals which is related to cybersecurity'
- **How**? Via (carefully) designed behavioural interventions
- **Aim**? Minimise the human attack surface

Security awareness training VS behaviour change

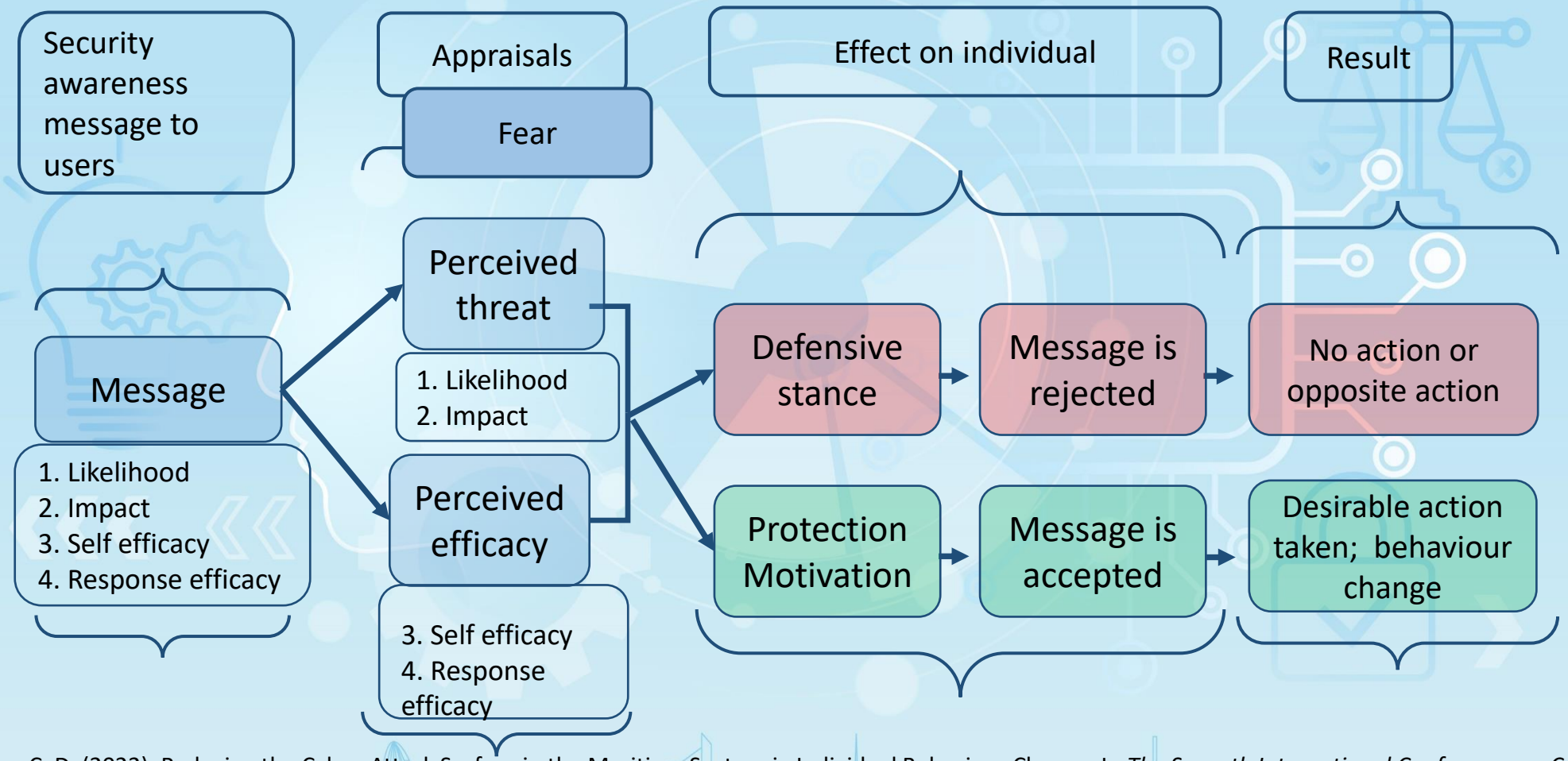
Awareness	VS	Behaviour
speculative		observable
indirect measurement, impact		Direct, objective measurement, impact
of secondary importance		priority, what matters

Types of behavioural interventions

- Fear appeals
- Nudges & boosts
- Conceptual frameworks
  - motivation, ability, easiness, triggers, rewards
- Nonconscious
- Incentives & disincentives

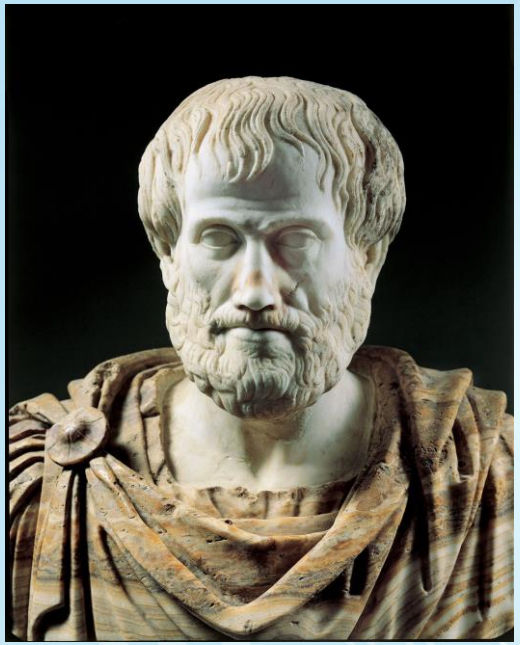
Behavioural interventions

Behavioural interventions (fear appeals)



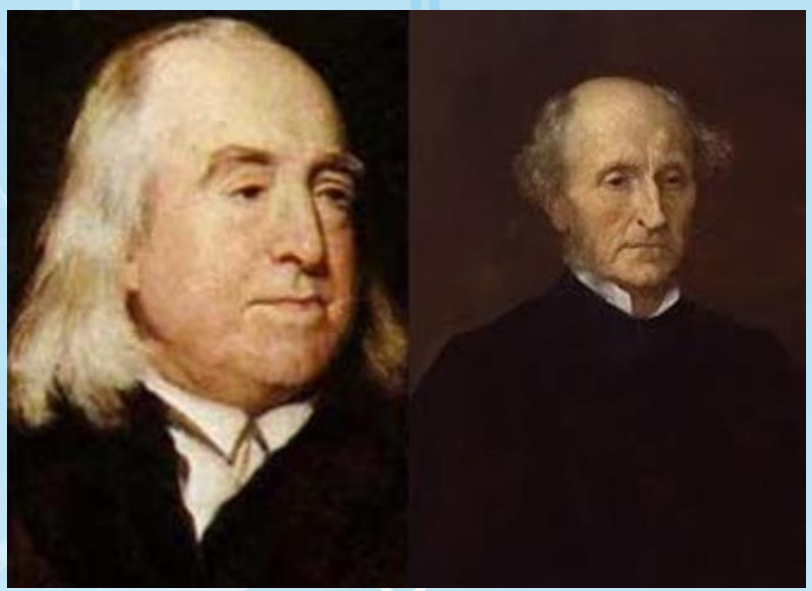
Mersinas, K., & Chana, C. D. (2022). Reducing the Cyber-Attack Surface in the Maritime Sector via Individual Behaviour Change. In *The Seventh International Conference on Cyber-Technologies and Cyber-Systems*.

Ethical traditions



Aristotle:
Virtue ethics

Immanuel Kant:
Deontological ethics



Jeremy Bentham &
John Stuart Mill:
Utilitarian ethics

Ethical principles

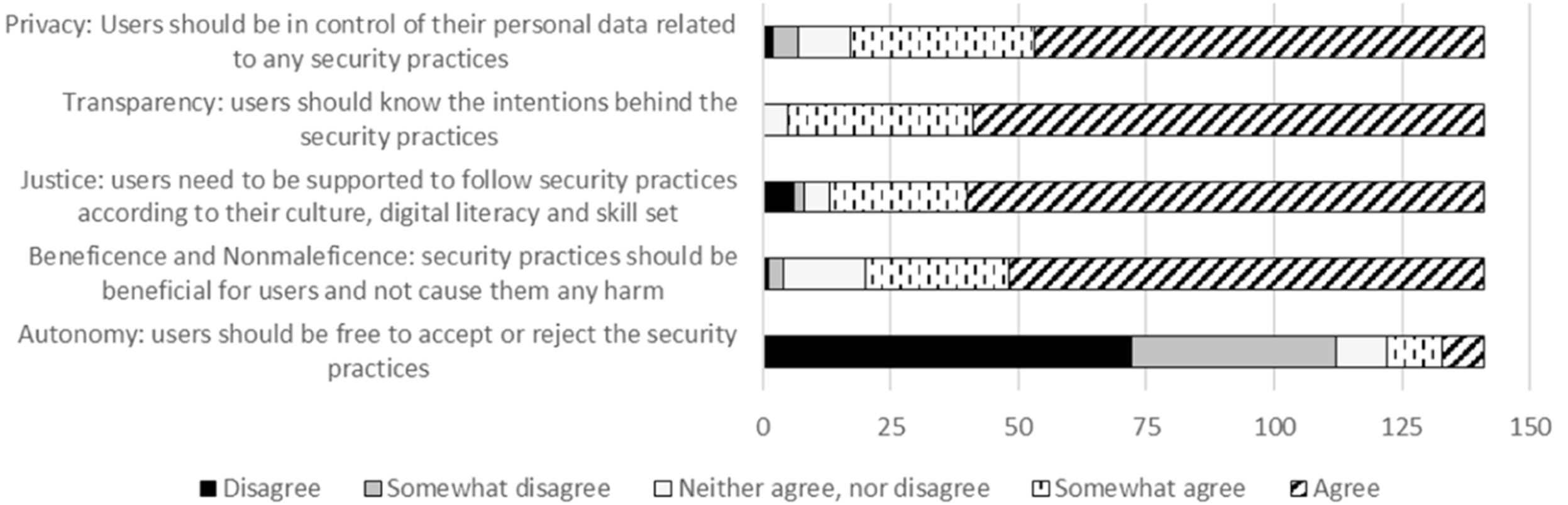
- autonomy
- nonmaleficence
- beneficence
- fairness
- justice
- privacy
- trust
- data protection
- data integrity
- consent
- equality
- transparency
- availability
- accountability
- accessibility
- confidentiality
- responsibility
- ownership
- usability

Six ethical principles

- Principle 1 - Autonomy
- Principles 2 & 3 - Beneficence & Nonmaleficence
- Principle 4 - Justice
- Principle 5 - Transparency
- Principle 6 - Privacy

Mersinas, K., Bada, M., & Furnell, S. (2025). Cybersecurity behavior change: A conceptualization of ethical principles for behavioral interventions. *Computers & Security*, *148*, 104025.

'The extent to which security professionals agree with the ethical principles' (N=141)



Behaviour change intervention steps

PROBLEM
Identify and define the problem, the actors, the desired outcome

DESIGN & CREATION
Craft the behavioural intervention

IMPACT / EFFECT
Measure the impact of the intervention considering direct and indirect effects

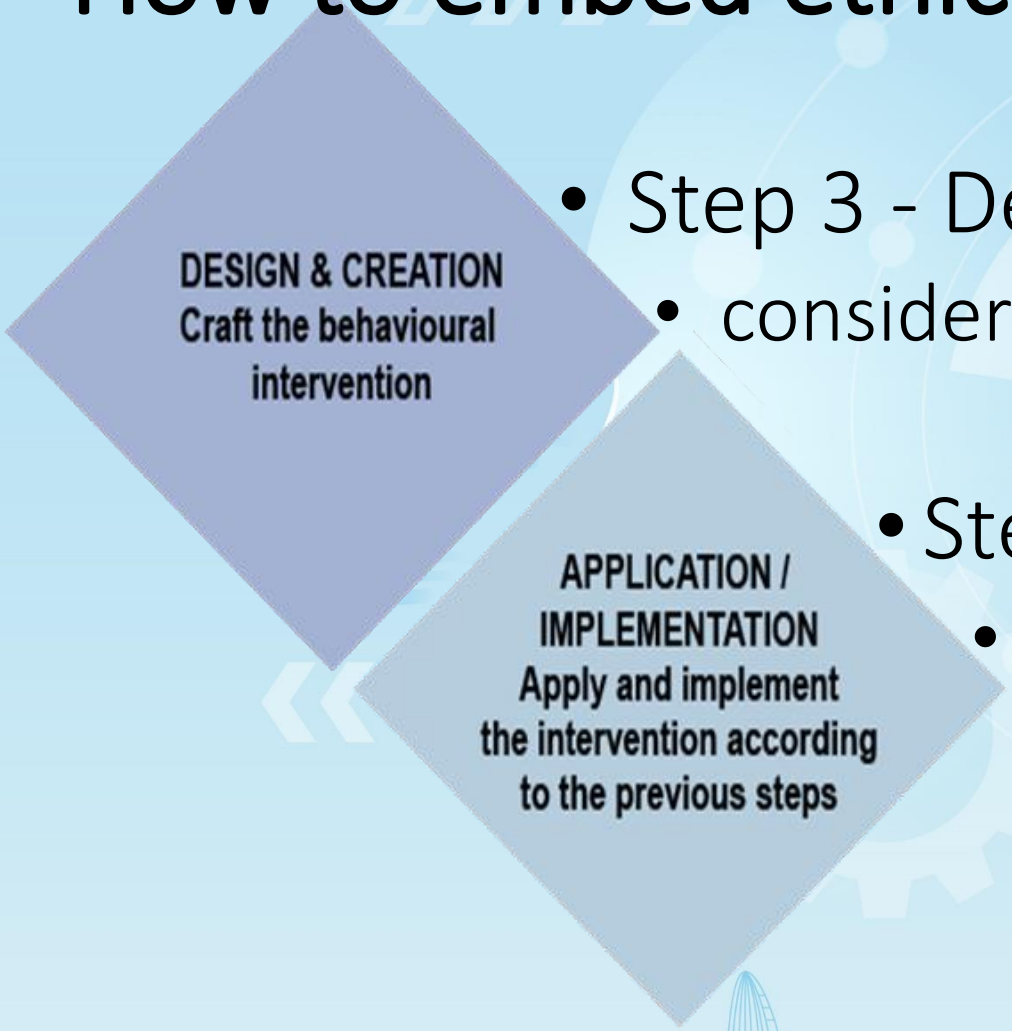
CONTEXT / ENVIRONMENT
Explore the context, the environment within which the intervention will take place

APPLICATION / IMPLEMENTATION
Apply and implement the intervention according to the previous steps

EVALUATION
Evaluate the whole intervention and decide the appropriate follow-up actions

Martin, K., Happa, J., Schmitz, G., Mersinas, K. (2025) *Cyber security foundations,* KoganPage, London.

How to embed ethics within the 6 steps?



- Step 3 - Design & Creation
  - consider ethics
- Step 4 – Application / Implementation
  - materialise ethics

AI-driven solution for behaviour change

cerc 2024
THE 6TH CYBERSECURITY EDUCATION & RESEARCH CONFERENCE

- What: Understanding human risk and responding in a customised fashion
- Aim: User behaviour change
- Functionalities:
  - Policy scanning, summarising, and explaining
  - Behavioural analysis of user activities → behaviour patterns
  - Categorisation of user activities and anomaly detection in behavioural patterns → indicative of security threats
  - Sentiment analysis in communication logs → interpret user intent and emotional indicators that could signify risky behaviours
  - Text mining and entity recognition → understanding user interactions with systems
  - Contextual AI, utilising contextual problems to adjust behaviour change guidelines dynamically based on real-time user activity data and threat landscape insights

المركز الوطني للأمن السيبراني
National Cyber Security Center

British Embassy Kuwait

KUWAIT COLLEGE OF SCIENCE & TECHNOLOGY
كلية الكويت للعلوم والتكنولوجيا

جامعة الكويت
KUWAIT UNIVERSITY

UK Science & Innovation Network

125 عاماً من الشراكة الكويتية البريطانية
125 YEARS OF KUWAITI-BRITISH PARTNERSHIP

Applying ethics to an AI solution (simplified)

Conceptual apparatus for security professionals, practitioners, designers

- Autonomy: users can reject or ignore the intervention ✓
- Beneficence & Nonmaleficence: by definition ✓
- Justice: customisation and adaptability ✓
- Transparency: processes and intentions ✓
- Privacy: information only provided to the user ✓

Thank you!

Dr Konstantinos Mersinas
Information Security Group, Royal Holloway, University of London

konstantinos.mersinas@rhul.ac.uk

*linkedin.com/in/**kmersinas***