AI Security Webinar

Practical Measures to Mitigate AI and Cyber Security Risks

11 June 2025

Innovate UK

BridgeAI

Agenda

11:10 Introduction to AI Security by Rosie Christos from Trilateral Research

11:30 Optimising Cybersecurity Investment: Reducing Component Vulnerability and Security Breach Risk by Richard Allmendinger, ISA for BridgeAI

11:40 Navigating security risks: SME perspectives, by Ed Blake from Absolute Security

11:50 Participants input: A live survey - poll

11:55 Panel Discussion and Q&A from the audience

12:20 Panelists final remarks

12:30 Close

UKRI Innovate UK BridgeAI

We encourage active participation

- Share thoughts on the chat
- As questions in the Q&A + upvote on questions you want to see answered
- Raise hand after the presentation
- Poll and feedback
- After the session, message us on skills@turing.ac.uk

Code of conduct

We are confident that our community members will, together, build a supportive and collaborative atmosphere at our events and during online communications. The following bullet points set out what we hope you will consider to be appropriate community guidelines:

- Be respectful of different viewpoints and experiences.
- Use welcoming and inclusive language.
- Do not harass people.
- Respect the privacy and safety of others.
- Be considerate of others' participation.
- Don't be a bystander.

As an overriding general rule, please be intentional in your actions and humble in your mistakes.

Adapted from & abiding to The Turing Way code of conduct available here.

Introduction to AI Security

UKRI Innovate UK BridgeAI

Meet your presenter

Rachel Finn, Trilateral Research

Rachel Finn is the Director of the Managed Services Division at Trilateral Research, which includes consultancy services in Data Protection, Cybersecurity and Responsible AI. She has handled large-scale information security and governance, risk and compliance projects for public and private sector organisations from international organisations like the UN to SMEs in the technology sector in the UK. Rachel has more than 15 years' experience designing practical solutions for information security, data protection and other human rights risks, enabling both innovation and protection. In addition, Rachel has co-developed methods for translating privacy, data protection and ethical requirements into technical requirements to support responsible and ethical AI. These have been used in many innovation projects, including Trilateral's internal software development. Rachel has a PhD from the University of Manchester (UK) and is widely published in the area of privacy, innovative data practices and new technologies.

Who We Are

TRILATERAL RESEARCH
Ethical AI

Global leader in Responsible AI (RAI)

20+ years' experience

Pioneering AI development

Multi-award-winning solutions and services

100+ clients (commercial and government)

Partnerships with tech/data regulators in 40+ countries

Headquartered in London (UK) and Waterford (IE)

Cybersecurity and AI

TRILATERAL RESEARCH
Ethical AI

78% of organisations are using AI in at least one business function. (McKinsey 2025)

1265% increase in phishing attacks since the launch of ChatGPT. (SlashNext, 2023)

47% of businesses using AI had no AI-specific cybersecurity measures in place. (DSIT UK, 2024)

Impacts of AI on Cybersecurity

TRILATERAL RESEARCH
Ethical AI

- AI-powered phishing
- Deepfakes for impersonation
- Data privacy breaches
- Automated vulnerability scanning
- Authentication disruption (CAPTCHA solving and password prediction)

LLMs reduce the cost of phishing attacks by **95%.**

(Harvard Business Review 2024.) https://hbr.org/2024/05/ai-will-increase-the-quantity-and-quality-of-phishing-scams

Cybersecurity risks for AI systems

TRILATERAL RESEARCH
Ethical AI

Data Poisoning

Manipulation of training data to force AI systems to produce inaccurate outputs.

Prompt injection

Entering a prompt into an LLM to enable a user to perform an unauthorised action

Data Privacy / Confidentiality

RAG systems that combine LLMs with use-case / organisation-specific data can expose personal data or proprietary, corporate confidential data to a new attack surface.

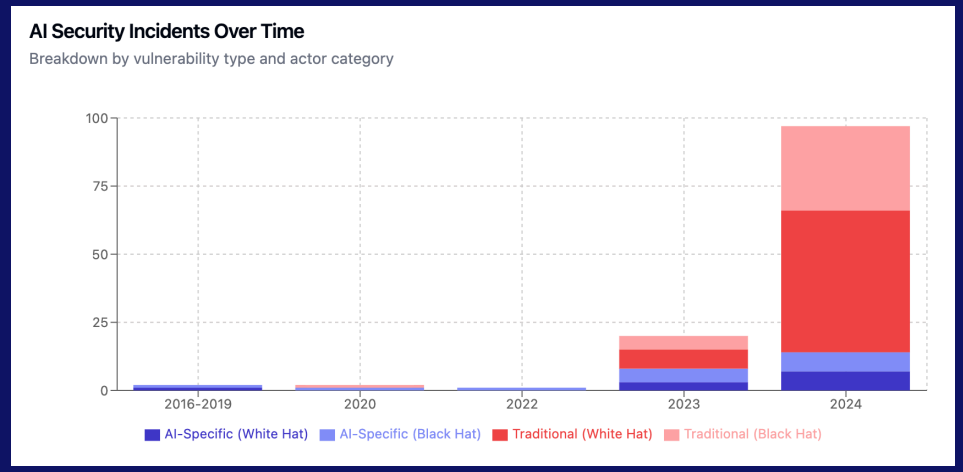
IT infrastructure weaknesses

Weakness of the infrastructure in which AI systems are deployed.

Scale of the threat

TRILATERAL RESEARCH
Ethical AI

AI-specific cybersecurity attacks are a relatively small proportion (18%) of overall attacks



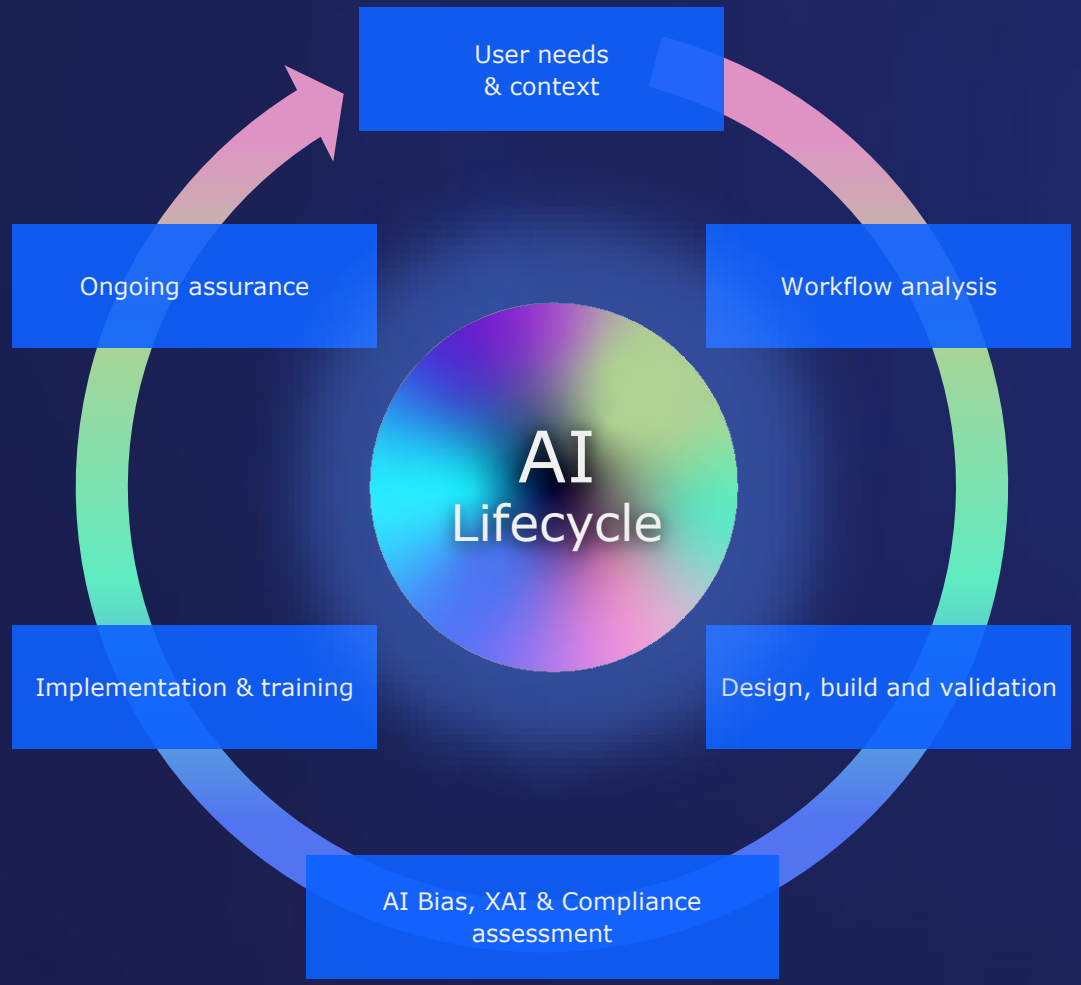
Sima, 2024
https://www.linkedin.com/pulse/real-story-behind-ai-security-incidents-caleb-sima-ge4yc/

The real danger is "conventional security failures that just happen to affect companies and software working with AI."

Sima, 2024
https://www.linkedin.com/pulse/real-story-behind-ai-security-incidents-caleb-sima-ge4yc/

Methods for securing AI systems & infrastructure

TRILATERAL RESEARCH
Ethical AI



- Infrastructure protection for training data and model deployment
- Data protection by design and default
- Cloud AI service security API protection
- Human oversight / human layer

Methods for securing software tools

TRILATERAL RESEARCH
Ethical AI

Action Plan

Align with relevant standards

- ISO 27000x / 42001, NCSC, CyberEssentials / +

Define

- Access control and authentication
- Acceptable use protocols

Implement

- Application monitoring and regular updates
- Network security and other system hardening protocols
- Vulnerability scanning and penetration testing
- Regular audits, risk assessments & security reviews

Benefits

Development of a cybersecurity culture by default

Strong risk management and response protocols

Strong third-party management and infrastructure protection

Staff awareness and training

- Cybersecurity, including strong awareness of social engineering
- AI Literacy, including real-world examples of AI misuse
- Data protection and confidentiality, including secure data handling
- Incident management, including incident reporting, tabletop exercises and simulations
  - Especially Phishing Simulations

TRILATERAL RESEARCH
Ethical AI

Regulatory Compliance in the UK

TRILATERAL RESEARCH
Ethical AI

Compliance builds trust and reduces legal risk—essential for SMEs using AI.

"Cyber security is an *essential precondition* for the safety of AI systems and is required to ensure, amongst other things, privacy, reliability, and the secure use of models."

UK Department for Science, Innovation and Technology (Jan. 2025)

Responsible AI outcomes

TRILATERAL RESEARCH
Ethical AI

- ✓ Better insights and information
- ✓ Confidence in deploying AI at scale across the organisation
- ✓ Confidence in regulatory compliance
- ✓ Greater customer trust
- ✓ Advancing of positive values in AI tools

Companies with a comprehensive, responsible approach to AI earn

2X as much profit

from their AI efforts.
(Bain & Co. 2024)

50% of responsible AI leaders say they have

better products and services than they would otherwise.
(BCG, 2022)

TRILATERAL RESEARCH
Ethical AI

Thank you!

Rachel Finn
Director - Managed Services
rachel.finn@trilateralresearch.com

Trilateral Research
London, UK | Waterford, IE

Cybersecurity Investment Optimization

UKRI Innovate UK BridgeAI

Meet your presenter

Richard Allmendinger, ISA for BridgeAI

Richard is Professor of Applied Artificial Intelligence, and Associate Dean for Business Engagement, Civic & Cultural Partnerships in the Faculty of Humanities, The University of Manchester, UK. He is also Co-Founder, COO and Head of AI of VeriBee Ltd, a UoM AI Cybersecurity spinout.

Richard is an Independent Scientific Advisor (since 2025) within the Innovate UK BridgeAI programme at the Alan Turing Institute, an Alan Turing Fellow Alumni, a Senior Scientist at Eharo, and AI Advisor for various organisations in private equity, BioTech, and EdTech.

UK Research and Innovation

SOTERIA PROJECT

MANCHESTER 1824
The University of Manchester

VeriBee

Innovate UK BridgeAI

Content

Background – Soteria project

Problem description

Stage 1: Delphi study

Stage 2: Optimization

Practical implications

Cybersecurity: Practice of protecting hardware and software from cyberattacks, unauthorised access, theft, or damage
Cyberattack: An attempt by hackers to damage or destroy a computer network or system

Soteria Project

- £5.8 million in funding from UKRI under the Digital Security by Design (DSbD) programme
- Project partners

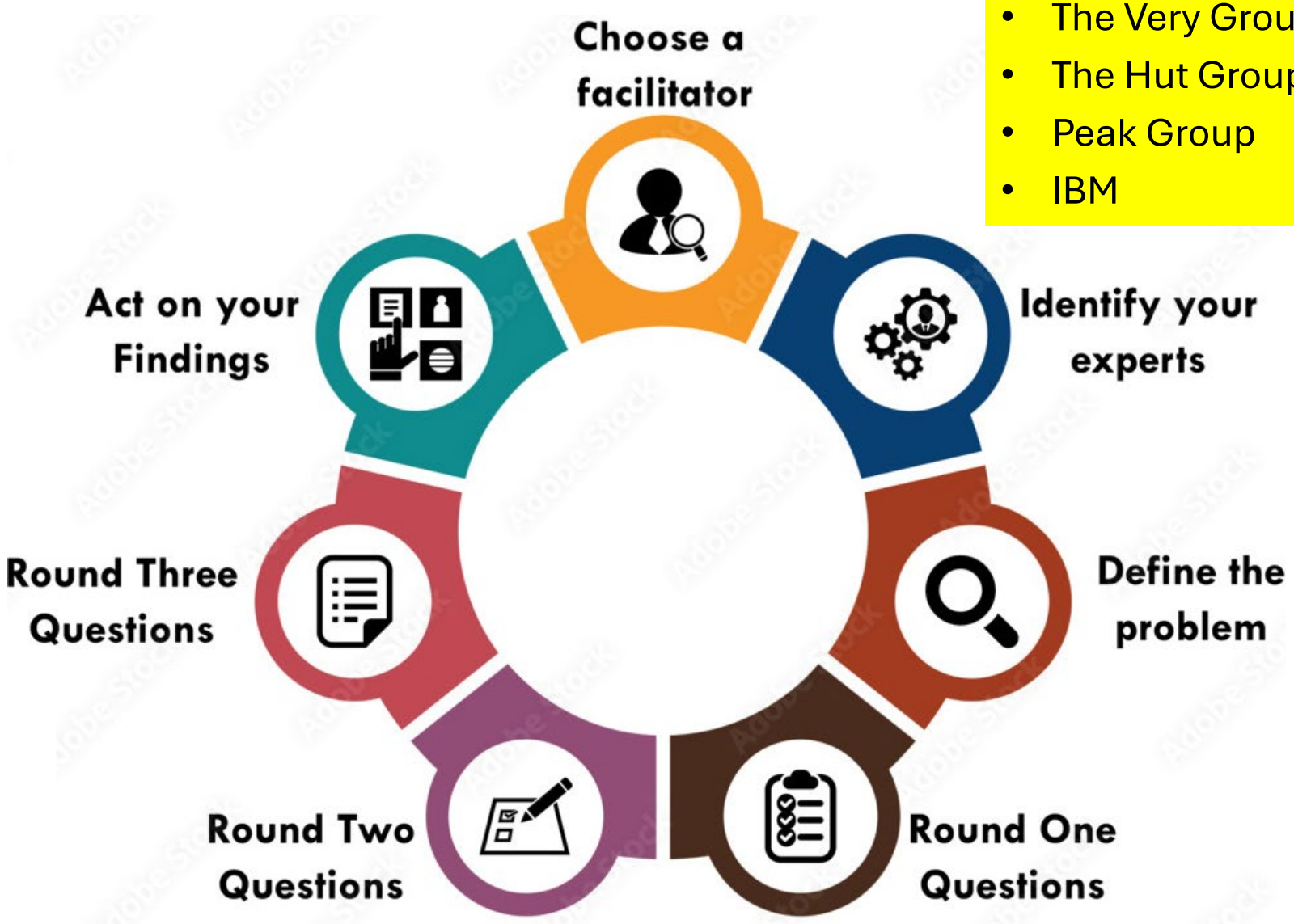
THG

MANCHESTER 1824
The University of Manchester

UNIVERSITY OF OXFORD

- Objectives of Business Optimization work:
  - Delphi study → identify most damaging types of cyber attacks and effective mitigation strategies
  - Investment optimisation → support decision-making in cybersecurity investment whilst reducing risk

Stage 1 – Delphi study



14 cybersecurity experts

- Microsoft
- The Very Group
- The Hut Group
- Peak Group
- IBM
- Co-op
- NCC Group
- City Football
- William Hill
- Cyber Smart
- The University of Manchester

Problem: Identify the most damaging types of cyberattacks and the most effective cybersecurity solutions

Stage 1 – Delphi study results I

Experts reached consensus on the **9** most damaging types of cyberattacks

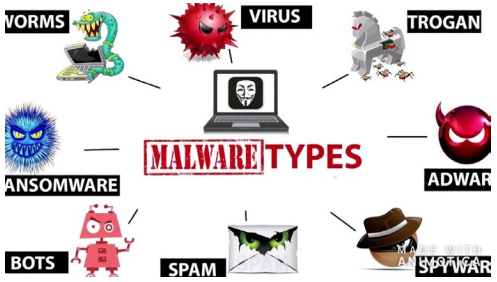
Ransomware attacks

Hacking

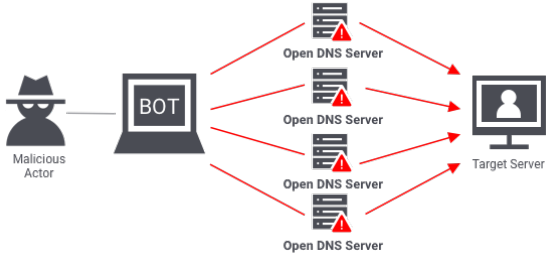
Take down of website

Phishing attacks

Social engineering



Viruses, malware, spyware

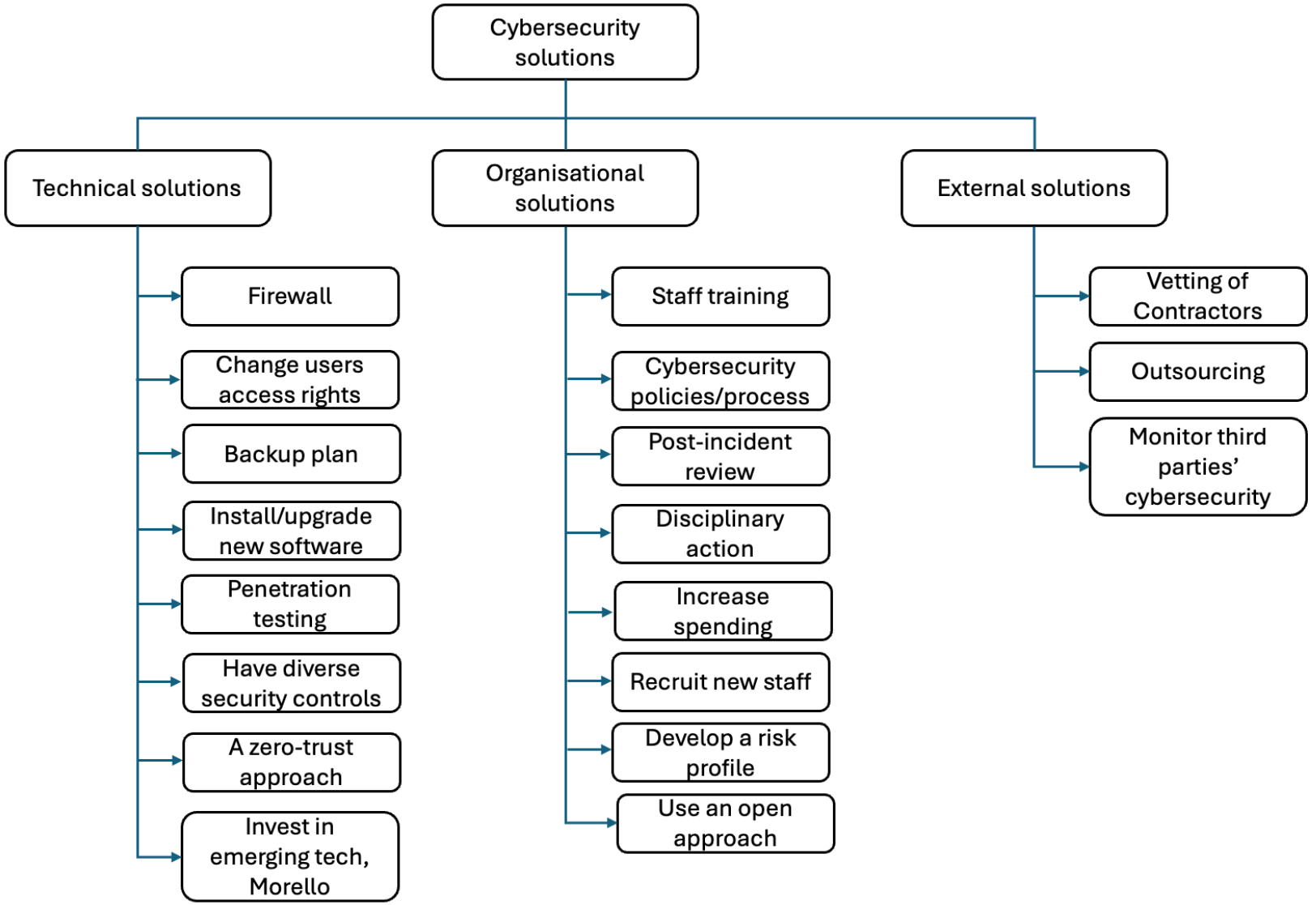


DDoS attacks

Impersonation

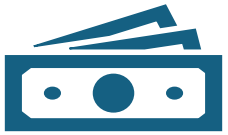
Stage 1 – Delphi study results II

Experts reached consensus on the **19** most effective cybersecurity solutions (**controls**)



Stage 2 – Cybersecurity investment problem

Given several critical **components** (hardware, software, human resources…) that need protecting, and potential **threats**, decide which cybersecurity **controls** to invest in such that we



Minimize cybersecurity cost (Implementation cost + expected cost of losses)

Minimize cybersecurity risk (Likelihood * impact)

Minimize environmental impact (CO2 emission)

subject to

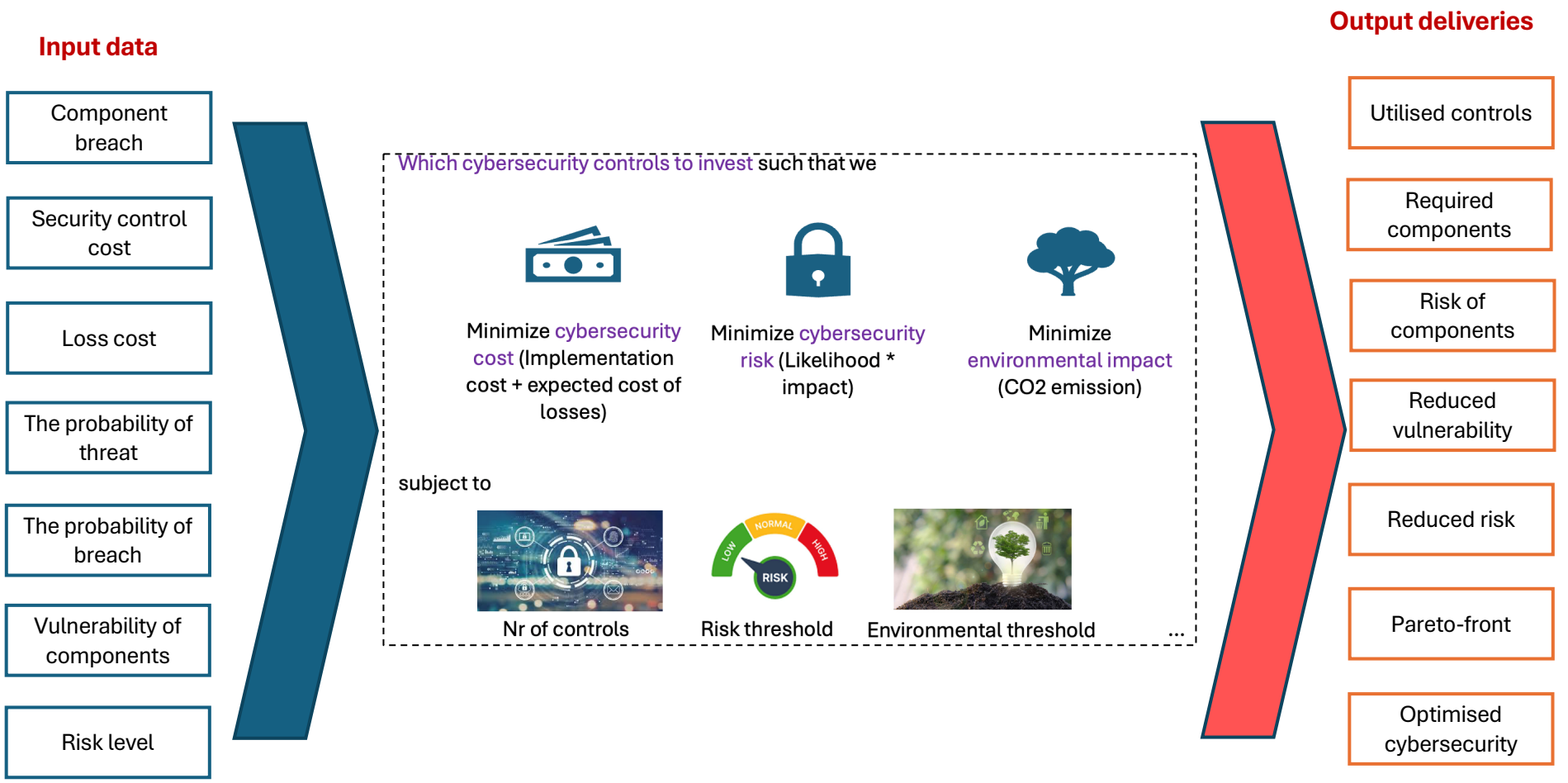
Nr of controls

Risk threshold

Environmental threshold

…

Stage 2 – Simulation framework

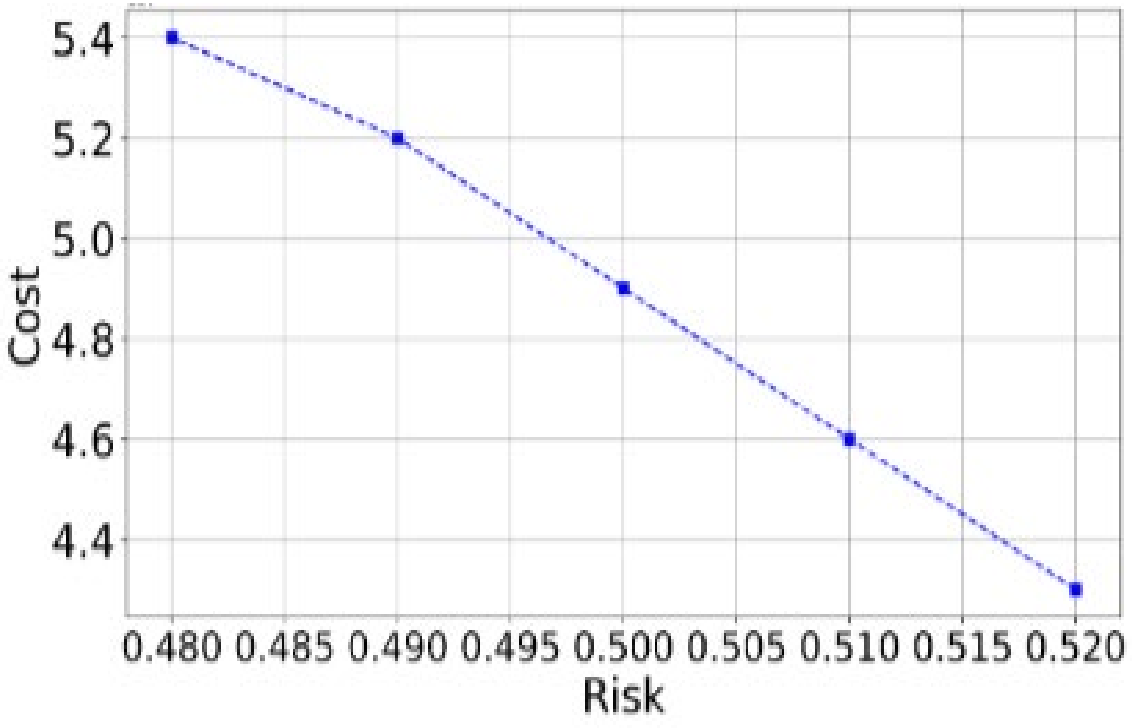


Stage 2 – Optimisation input data

1	**Threats** hacking, DDoS, Virus, ransomware, …	1) Identify **9** cybersecurity threats from the **Delphi study**. 2) Provide the probability of occurrence for each threat within their business cycle. 3) Input threat-component relationship.
2	**Security Controls** firewall, penetration testing, outsourcing..	1) Identify **19** cybersecurity controls from **Delphi study**. 2) Input the implementation costs. 3) Specify the efficiency coefficients for each control-component combination.
3	**Components** network, servers, data, users…	1) Identify **5** critical components of the organization's IT infrastructure. 2) Input intrinsic vulnerability of each component without any security controls. 3) Input the cost of potential losses associated with a security breach of a component.
4	**Risk**	1) Rank likelihood of threat on a scale of 1 (Rare) to 5 (Highly Likely), and its impact on a scale of 1 (Negligible) to 5 (Very Severe)
5	**Sustainability**	1) Rank the impact of each control on CO2 emission on the scale of 1-5. 2) Rank control-component sustainability relationship on a scale of 1-5.

Data source: Cybersecurity Director of THG and Chief Information Security Officer of UoM

4. Stage 2: Optimization results I



Reducing cybersecurity risk is expensive…

4. Stage 2: Optimization results II

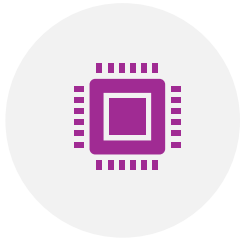
Control ID	Security controls description (from Delphi study)	UoM 3 Obj Optimum	UoM 2 Obj Optimum	THG 3 Obj Optimum	THG 2 Obj Optimum
1	Firewall	**0**	1	**0**	1
2	Antivirus/anti-malware software	**0**	1	**0**	1
3	Data Encryption	**0**	0	**0**	1
4	Backup plans and solutions	**0**	0	**0**	0
5	Principle of Least Privilege	**0**	0	**1**	1
6	Penetration testing	**0**	0	**1**	1
7	Invest in Morello	**1**	0	**1**	0
8	Staff training	**1**	1	**1**	1
9	Recruit new staff	**1**	1	**1**	1
10	Additional vetting of staff or contractors	**1**	1	**1**	1
11	Cybersecurity policies/procedures	**1**	1	**1**	1
12	Disciplinary action	**1**	1	**1**	1
13	Post-incident review	**1**	1	**1**	1
14	Monitoring third parties' cyber security	**1**	1	**1**	1
15	Outsource cybersecurity	**1**	1	**1**	0
16	Develop a risk profile	**1**	1	**1**	1
17	An open approach	**1**	1	**1**	1
18	A zero-trust approach	**1**	1	**1**	1
19	Cybersecurity insurance	**1**	1	**1**	1

Differences between UoM and THG due to varying probabilities of attack, vulnerabilities of components, loss cost in case of security breach, implementation costs, risk level.

Practical implications

Smart Trade-offs

You can’t eliminate all cyber risks—but our model helps you strike the right balance between risk, cost, and environmental impact, tailored to your systems.



Higher Risk, Higher Cost

Reducing cyber risk—especially in high-threat environments—requires more investment in security controls.

Vulnerability Drives Cost

The more vulnerable your systems, the more you’ll need to spend. Building secure systems early reduces long-term cybersecurity costs.

Informed Investment Choices

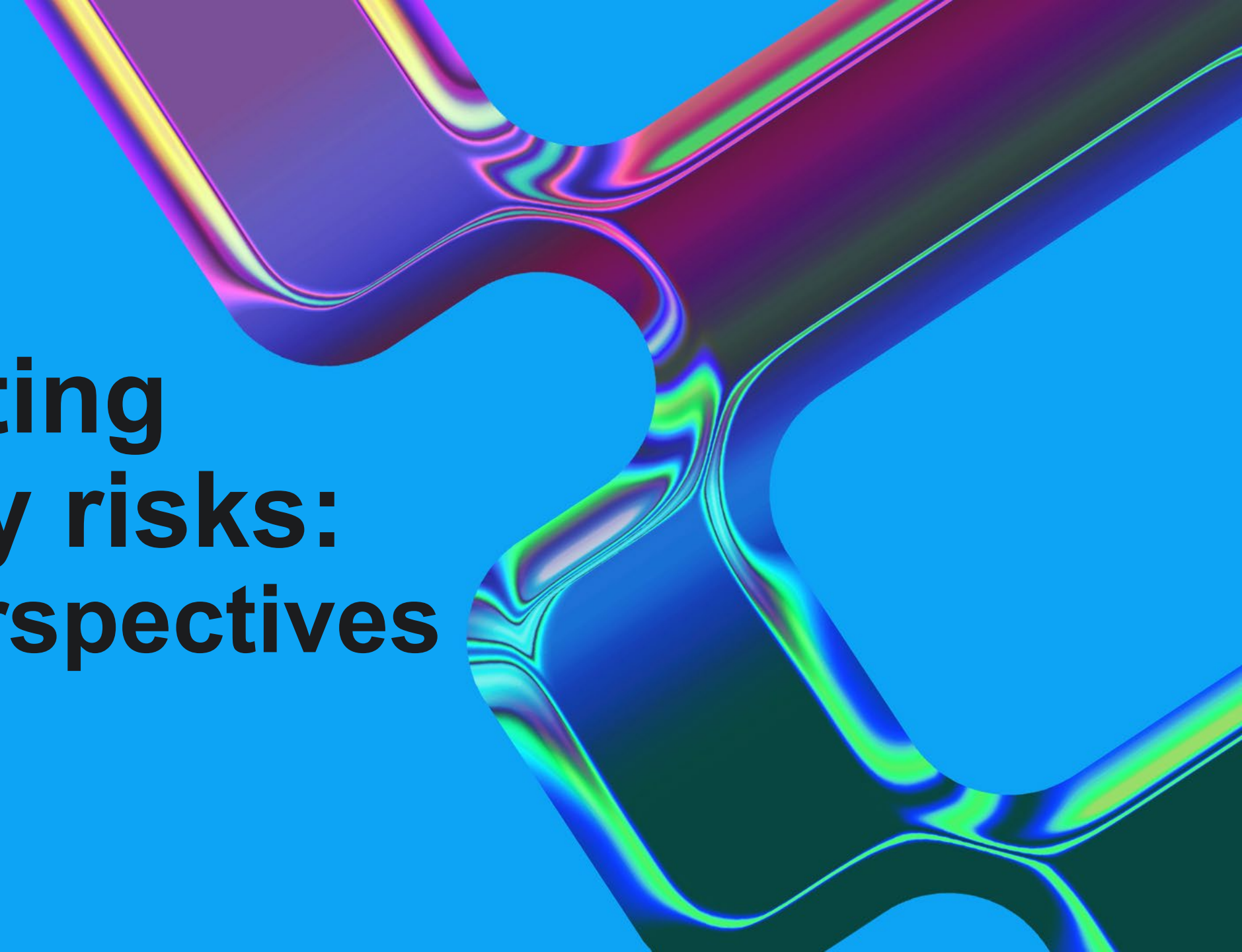
A suitable optimization method can give you a menu of cost–risk trade-offs, so you can invest based on your risk appetite and budget.

MANCHESTER 1824
The University of Manchester

Thank you!

richard.allmendinger@manchester.ac.uk

Navigating security risks: SME perspectives



UKRI Innovate UK BridgeAI

Meet your presenter

Ed Blake, Absolute Security

Ed Blake has a BSc in Geographical information Systems covering Geospatial systems , Artificial intelligence and Remote Sensing. He has 29 years experience in Cyber Security solutions and has helped pioneer the market for Web Application Firewalls (WAF) and Security Orchestration Automation and Incident Response (SOAR) platforms. A regular Infosec presenter who has presented at the E-Crime Congress in London and Dubai, ISACA London, BAPCO, PCI London, Gartner Security and risk Management, Barracuda Networks European Partner conference and a number of IBM hosted events and has been quoted regularly in the UK IT and National press over the last 25 years. He has worked on some of the worlds largest Cyber Security and Financial IT projects.

ABSOLUTE SECURITY

Our unique enterprise resilience platform – rooted in hardware – minimizes the risk and dramatically mitigates the impact of business disruptions

650M
Absolute embedded devices

3.1B
Daily API calls from endpoints

227
Patents worldwide

28
Global PC OEMs

UKRI Innovate UK BridgeAI

RANSOMWARE IS ON THE RISE

RANSOMWARE REMAINS THE LARGEST IT DISRUPTOR

↑104%

The number of Ransomware attacks more than doubled in 2023

75%

Most companies had an event in 2023

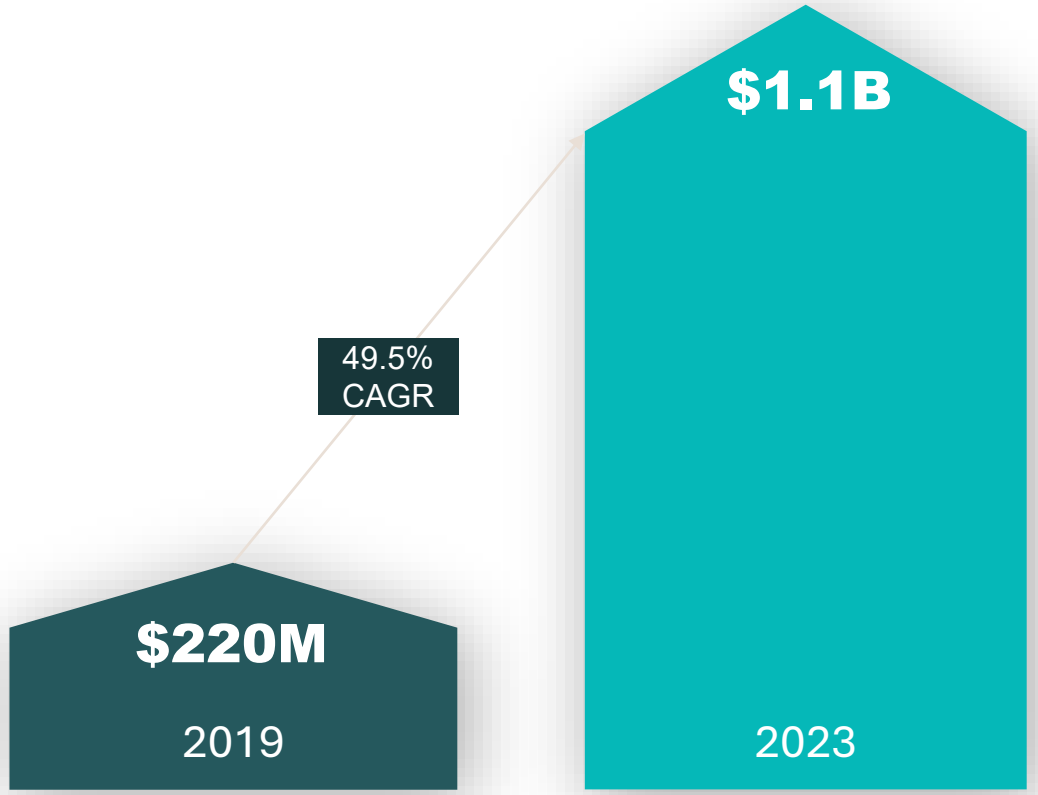
43%

Of affected data is not able to be restored

32%

Ransom is only a third of the financial impact

Total $USD Value Received by Attackers



Innovate UK BridgeAI

Sources: (1) Armis: The Anatomy of Cybersecurity: A Dissection of 2023's Attack Landscape (2) Gigamon + Commvault Cyber Recovery Readiness Report 2024; (3) Chainalysis Crypto Crime Report 2024; (4) Veeam Ransomware Trends

Customer Situation – Retail - Ransomware

- The customer’s entire business was halted by Hard2Decrypt ransomware
- Trusted security tools rendered inoperable – compromised by Ransomware
- Customer war room, security vendors and consultants at an impasse
- Security operations director called our Absolute Assist technical consultant
- Absolute Assist consultant team joins war room

Ransomware response timeline

Week 01

- Phishing attack occurred at some point prior
- **October 31st** Ransomware detected. CiscoAMP compromised.
  -War room created.
- **November** attempts to deploy SentinelOne fail due to Ransomware corruption
- **November 6th** Absolute Engaged.
- **PS Team restored SentinelOne & Cisco AMP through Absolute's Application Persistence.**

Week 02

- **November 7-14:**
- **Custom "Reach" scripts** deploy triage tools
- **Device Freeze** applied to **all** systems to halt user interaction and inform users of attack.
- **Custom PS Reach script** used to detect scope of enterprise-wide breach. Identify "Safe" vs "Unsafe" devices.
- SentinelOne "quarantines" all systems.

Week 03

- **Nov 15-30:** Recovery begins.
- **Custom Reach scripts** used to deploy local accounts to all endpoints.
- **Custom reach scripts** used to deploy additional security tools.
- **Recovery workflows created**. Custom Reports, end user messaging, application health, scripts.
- Systems begin coming back online.

Week 04

- **Dec 1- onward**.
- Compromised systems begin reimaging and restoration.
- Audit team continual review of security and status reports from Absolute.
- Absolute Assist team on standby to assist with ongoing recovery efforts.

Absolute Capability	Purpose	Result
Absolute Core Firmware Agent Persistence	Tamper resistant foundation to Absolute services	Provide recoverability in the event of attack which drives healing up the Persistence Chain
Application Persistence	Monitor core applications (such as APT) health, report on compliance and repair when unhealthy.	Core ransomware and security applications are always ready when needed to help prevent and to respond to an attack. Anti-malware apps like SentinelOne and CiscoAMP restored to functionality within minutes.
Absolute Reach Scripting Platform	Provide a platform for remote device actions that is flexible and resilient.	- Backup installation of applications (SentinelOne, Bomgar, CiscoAMP) and custom JIT deployments. - Modify system states (reboots, registry mods, tool execution) - Perform arbitrary data gathering (ie scan for encrypted files) - Recover user access to devices (add local admin accounts)
Absolute Containment/quarantine	Restrict device network access to halt the spread of a malicious agent	- Inbound and Outbound network traffic is halted - Absolute services remain accessible to the Absolute agent - Additional services can be allowed based on whitelist - API or console driven to lockdown target devices.
Absolute Assist - Consulting	Provide process insights prior to and during a ransomware attack.	Help customers prepare a resilient endpoint state to help protect against an attack as well as strategies and mechanisms to respond if an attack occurs.
End User Messaging	Inform users of incident status, provide instruction.	Compromised devices may need to be locked down or otherwise restricted. User access may need curtailed. User awareness is critical to a well-managed response.
Device Freeze	Lock down devices to prevent user actions and inform.	Preventing user interaction with infected systems is critical.
Absolute Custom Device Fields		Detection and Response actions may generate output that is useful in reporting and analytics.

Script description	Purpose	Limitations
Add Local Admin Account(s) to endpoints	Inject back local admin account to recover access for users to systems. Domain access was disrupted due to active directory domain controller infections.	
Force Reboot script	Reboots were required for several application installations and self-healing of the Absolute agent.	
Install Bomgar – remote access tool	Script to download and install the Bomgar agent on target systems. Bomgar was then able to provide remote control access to these systems for further remediation efforts.	File hosting was a challenge for the customer. Absolute PS provided hosting.
File Detection Script – compromised files	Search for an arbitrary file pattern, recursively through a specified root path (ie c:\) and return a "File Found" or "File Not Found" result. Results were stored in a CDF for use in reporting (filtering, sorting and targeting of devices based on state).	Anti-malware solutions could NOT detect static encrypted files but Absolute's script could.
Backup installation scripts for SentinelOne and CiscoAmp	As a backup method of delivery scripts to download and install these applications were created. Application Persistence negated the need for these.	File hosting was a challenge. Absolute PS hosted.

Customer Situation – Healthcare - Ransomware

- The customer’s entire business was halted by Ransomware – Phone Systems, Healthcare Records, Patient treatment data
- Trusted security tools rendered inoperable Again! – compromised by Ransomware
- 2 Million Patients affected in coverage area.
- Customer war room – Anti-Malware technology missed dormant encrypted files on devices (No C&C Traffic)
- Security operations director called our Absolute Assist technical consultant
- Absolute Assist consultant team joins war room

Application Persistence (AP) – the Foundation of a Resilient Security Platform

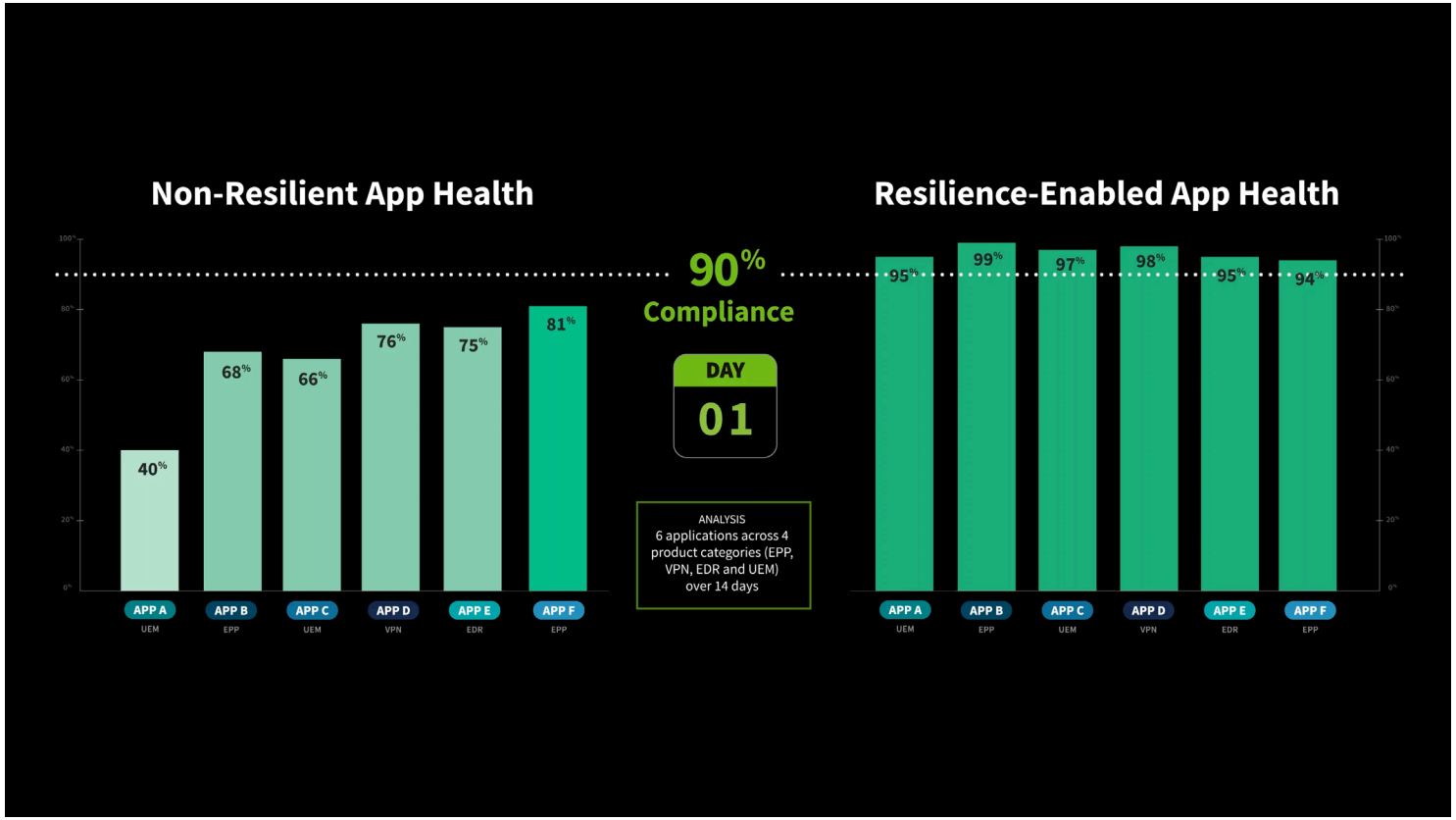
Deploy, Monitor and Heal Core Security Applications

- **Ransomware attacks** consist of a **multi-stage attack system** to both compromise and weaken the ability of security applications to respond.
  - Application corruption – encryption of application files –Anti-Malware was corrupted during deployment
  - Application removal – Ransomware uninstalled existing security controls using a compromised uninstall password.
- AP for **Security Applications** should be on devices BEFORE an attack occurs.
- AP Reinstall can protect applications against compromise so that they can detect ransomware attacks early.
- AP can deploy and protect applications after an attack has occurred.

Cyber Resilience: Why It Matters (Example Applications)

Factors impeding security efficacy:

- Re-imaging of a device often leads to the agent not being re-installed.
- Critical files are often corrupted while new third-party apps are installed or updated.
- End users disable apps unintentionally or for uninterrupted work.
- Threat actors disable mission-critical apps to bypass security controls and avoid detection.



Driving the Need for Application Resilience

Innovate UK BridgeAI

THANK YOU

skills@turing.ac.uk