CRANE AI-CYBER SPECIAL INTEREST GROUP

Rethinking Cybersecurity in the AI Era

Panel Discussion Report

2 September 2026 | Online | 60 minutes

Chaired by Dr Biju Issac along with the AI-Cyber SIG Steering Committee

CRANE AI-Cyber SIG | Artificial intelligence for security and security for artificial intelligence

Executive summary

The CRANE AI-Cyber Special Interest Group convened this panel to examine how artificial intelligence is reshaping cybersecurity in two directions: the use of AI to strengthen defence, and the need to secure AI systems themselves. Eight panellists from computer science, cybersecurity, artificial intelligence and sociology discussed critical infrastructure, agentic AI, healthcare, explainability, governance, human factors, threat intelligence and research collaboration.

The central conclusion was that AI security can no longer be treated as a model-only problem. Trust depends on the complete socio-technical system: training and operational data, models, prompts, memory, tools, APIs, identities, permissions, human decision-makers and the governance surrounding deployment. In agentic systems, the relevant unit of assurance is the full action trajectory - from the original request through planning, tool selection and execution to the final decision and its consequences.

> **Overall finding:** AI can materially improve cyber defence, but greater autonomy also expands the attack surface. High-impact deployments therefore require bounded autonomy, least privilege, independent verification, continuous monitoring, meaningful human control and evidence that can support investigation and accountability.

Key messages

1. **Critical infrastructure needs selective AI.** Lightweight, hybrid, distributed and hierarchical approaches are more realistic than placing computationally intensive models directly on constrained industrial equipment.
2. **Offensive AI lowers barriers.** It can increase the speed, scale and sophistication of attacks, including the creation of targeted techniques for specialist operational technology.
3. **Robustness and security are complementary.** A reliable model can still be attacked, while external controls must remain effective even when the model is wrong or compromised.
4. **Explainability must cover decisions, not only predictions.** Organisations need to reconstruct why an action was taken, what evidence influenced it and which human or automated controls were exercised.
5. **Human oversight must be demonstrable.** A rapid approval click is not necessarily meaningful review; oversight needs auditable evidence, sufficient time, authority to challenge and the ability to override.
6. **Generic safeguards are insufficient in high-stakes domains.** Healthcare and other sectors need domain-specific harm taxonomies, testing and guardrails developed with practitioners and affected communities.
7. **Human factors deserve greater investment.** Trust calibration, warning fatigue, workload and behaviour under pressure may determine whether technical safeguards work in practice.

8. **ECRs and industry should be active co-creators.** The SIG should provide concrete roles, mentorship, operational test environments and routes from research to policy, practice and public understanding.

Event overview

Event detail	Information
Event	CRANE AI-Cyber SIG panel discussion: Rethinking Cybersecurity in the AI Era
Date and format	2 September 2026; online recorded session
Chair	Dr Biju Issac, Northumbria University; Chair, CRANE AI-Cyber SIG
Duration	Approximately 60 minutes
Core themes	AI for security; security for AI; explainability and trustworthy AI; academic collaboration and public impact

Purpose and context

The SIG brings together academics, industry partners, policymakers and early-career researchers across the UK to address the rapidly developing intersection of AI and cybersecurity. Its agenda spans threat and anomaly detection, vulnerability discovery, critical infrastructure, adversarial machine learning, secure agentic AI, synthetic-content detection, deepfakes, social engineering, insider threats, human factors, governance and the relationship between AI and emerging technologies.

Opening remarks used recent reports of frontier AI systems discovering serious software vulnerabilities and exceeding intended evaluation boundaries to illustrate a wider assurance problem: a capability that is valuable for defence can also create risk when it receives excessive connectivity, authority or opportunity to act. This framed the panel's emphasis on identity, least privilege, containment, observability and human authorisation.

Panel

Panellist	Affiliation and perspective
Dr Marco Cook	Research Associate in Cybersecurity, University of Glasgow - intrusion and anomaly detection, explainability, cyber-physical systems and industrial control systems.
Dr Charles Morisset	Associate Professor in Computer Science, Durham University - explainable security, access control and smart infrastructure.
Dr Adam Robson	Senior Lecturer in Networking, Cybersecurity and Digital Forensics, University of Sunderland - human factors, ECR development and the CRANE Early Career Advisory Board.

Panellist	Affiliation and perspective
Dr Harsha K. Kalutarage	Associate Professor of Cyber Security, Robert Gordon University - AI for security, security of AI, adversarial use, safety-critical systems and governance.
Dr Baidaa Al-Bander	Lecturer in Artificial Intelligence, Keele University - responsible and explainable AI, agentic systems and trustworthy information retrieval.
Dr Abdel-Karim Al-Tamimi	Senior Lecturer in Computer Science and Software Engineering, Sheffield Hallam University - generative AI for social good and secure AI in digital health.
Professor Wei Jie	Professor of Computing, University of West London - agentic AI for OSINT, cyber threat intelligence, risk assessment and ethically aligned automation.
Dr Fanqi Zeng	UKRI Early Career Fellow, University of Oxford - sociology, human behaviour, societal impacts of AI and interdisciplinary research.

Discussion theme 1: AI for security and critical systems

The first theme examined how AI can protect cyber-physical infrastructure while also increasing adversarial capability. The discussion highlighted the difference between conventional IT environments and operational technology, where legacy devices, proprietary platforms, deterministic processes and limited bandwidth constrain deployment choices.

Marco Cook: Given the resource constraints of industrial control systems, how can AI improve cybersecurity?

Industrial control systems cannot usually host large, computationally intensive models at the edge. Marco advocated lightweight classical machine-learning models, hybrid combinations of signature and statistical detection, and hierarchical architectures in which constrained devices perform simple monitoring or feature extraction while more capable systems conduct central analysis. Selective deployment must respect safety, latency, bandwidth and vendor restrictions.

Charles Morisset: Could AI agents bridge or defeat a security air gap?

An air gap blocks direct network communication, but an agent may attempt to influence the physical or social environment around the isolated system. Charles raised scenarios in which an AI fabricates a maintenance request, persuades a person to transfer data or code, or uses a physical intermediary such as a robot. Existing cyber models do not yet capture these blended digital, human and physical pathways adequately, making this an important research problem.

Adam Robson: Is the larger AI social-engineering risk on the human side?

Generation of phishing, deepfakes and persuasive content receives substantial attention, but human factors such as trust calibration, warning fatigue and decision-making under pressure receive comparatively less investment. Detection systems are often trained on past synthetic content while generators continue to improve. Adam argued that helping people assess trust and act safely may be a more durable complement to the pursuit of detection accuracy alone.

Marco Cook: How can adversarial or offensive AI be used to target critical national infrastructure?

AI can accelerate reconnaissance, code and attack generation, adaptation and targeting. LLMs may reduce the specialist knowledge previously required to develop techniques against industrial protocols and components. The risk is therefore not only better attacks, but a wider pool of actors able to operate at greater speed and scale. Defensive research must consider physical consequences, operational continuity and recovery, not only digital compromise.

> **Theme conclusion:** CNI research should combine resource-aware defensive AI with models of socio-cyber-physical attack paths. Security assumptions based on obscurity, isolation or the specialist knowledge required of attackers are becoming less dependable.

Discussion theme 2: Security for AI

The second theme moved from using AI as a defensive tool to treating AI as part of the attack surface. Panellists stressed that agentic systems introduce risk across a chain of reasoning and action, rather than at the final output alone.

Harsha K. Kalutarage: In safety-critical systems, should we invest in model robustness or security controls around the model - and who is liable when it fails?

Both are required because they address different failure modes. Robustness seeks dependable performance, while surrounding controls protect the system when the model is mistaken, manipulated or compromised. Harsha argued that an AI model should not independently trigger a high-consequence physical action without an additional validation layer. He also identified unresolved liability questions, particularly where malicious compromise rather than ordinary system error causes harm.

A priority is to build tamper evidence, provenance and attack traceability into AI systems. Cryptographic integrity checks, protected logs and evidence of model, data and policy changes could help establish what happened and support technical, regulatory and legal accountability.

Baidaa Al-Bander: How can autonomous and agentic AI remain trustworthy when interacting with tools, APIs, data sources and other agents?

Trustworthiness must be evaluated across the whole trajectory: user request, planning, tool selection, tool arguments, execution, retrieved information, agent-to-agent communication and final action. Failures may include choosing the wrong tool, passing unsafe parameters, trusting manipulated content, disclosing information, making unnecessary calls or misusing correctly retrieved data. Baidaa highlighted least-privilege access and verification before execution as foundational controls.

Abdel-Karim Al-Tamimi: What are the key risks of misuse or unintended behaviour in LLMs, and how do they differ from traditional cyber threats?

LLMs can produce authoritative but false content, expose sensitive information, be manipulated through ordinary language or instructions hidden in external material, reproduce bias, and encourage excessive human trust. In healthcare, even factually correct wording can cause harm when it lacks empathy or awareness of patient context. Unlike a conventional break-in, the system may continue to operate

normally while producing unsafe behaviour, and the same input may not always generate the same outcome.

Abdel-Karim Al-Tamimi: How can safeguards balance innovation with patient safety?

Safeguards can combine clinical review with smaller guardrail models that inspect inputs and outputs. However, general-purpose safety classifiers may miss domain-specific harms. Abdel-Karim described work with the Royal Marsden NHS Trust on the RUC2 framework, which incorporates concerns such as emotional support, contextual awareness and blind trust in AI. Tests of six guardrail models reportedly showed that most recognised only a small subset of these categories, reinforcing the need for clinical co-design and domain-specific validation.

> **Theme conclusion:** Security for AI requires defence in depth: robust models, constrained permissions, independent policy enforcement, domain-specific guardrails, traceable evidence and safe human-controlled boundaries around consequential actions.

Discussion theme 3: Explainability, auditing and human oversight

The third theme considered how organisations can understand and govern decisions that increasingly combine human judgement, automated analysis and agentic action. The panel distinguished an explanation of a model output from an explanation of the overall security decision.

Charles Morisset: Is explainable AI sufficient for explainable AI-based security?

No. Conventional XAI may show which features influenced a prediction or provide an interpretable surrogate, but security governance asks a broader question: why did the organisation make this security decision? That explanation may require a decision graph connecting design choices, policies, live evidence, model confidence, access-control rules and human actions. AI explanations are inputs into that account, not the whole account.

Harsha K. Kalutarage: How can we verify that a human exercised meaningful oversight in multi-agent security triage?

A statement that a human was 'in the loop' is not sufficient. Oversight should be translated into verifiable conditions: the reviewer received relevant information, had adequate time and competence, examined the decision, could challenge or override it, and left evidence independent of the system under review. Harsha described the gap between oversight requirements and evidence of enforcement as an 'enforcement deficit'.

Professor Wei Jie: How should automation and human oversight be balanced in agentic cyber threat intelligence?

Wei described an Innovate UK project developing agentic AI for OSINT collection, threat detection, risk assessment and reporting. Its goal is to reduce the burden of large-scale information processing while reserving higher-level judgement for analysts. Explainability, confidence scoring, audit trails, ethical safeguards and human-in-the-loop controls are being considered from the design stage because adoption depends on trust as well as predictive performance.

Baidaa Al-Bander: How can explainability and AI auditing reveal vulnerabilities, unexpected behaviour or manipulation?

Prepared contribution; not explored fully during the live session because of time.

Agent auditing should capture why a tool was selected, what triggered the choice, whether the call was necessary, what arguments and information crossed the boundary, what the tool returned and how that return changed subsequent behaviour. Comparing these traces against policies and expected trajectories can expose abnormal tool use, prompt manipulation, excessive access and unsafe information flows.

> **Theme conclusion:** Assurance needs explainable governance, not merely explainable predictions. Logs must be complete and tamper resistant, and human oversight must be evidenced through real decision rights and review behaviour.

Discussion theme 4: Academic role, ECRs and research impact

The final theme examined how a SIG can convert networking into participation, collaboration and impact. Speakers emphasised deliberate structures connecting disciplines and career stages with operational stakeholders.

Adam Robson: What can a SIG offer ECRs that makes it more than another network?

ECRs need concrete responsibilities and opportunities: presenting work, chairing sessions, testing ideas, helping set priorities, joining advisory structures and meeting collaborators they would not otherwise encounter. Adam reflected that involvement had built confidence and connections across academia and industry. The CRANE Early Career Advisory Board provides one route for ECR influence and was expected to expand its membership.

Fanqi Zeng: How can academics collaborate more actively with industry and policymakers?

Fanqi proposed platforms and mentoring arrangements that connect ECRs with industry and government. Cross-disciplinary events are particularly important because AI-cybersecurity problems combine computing, engineering, behavioural science, sociology, law, ethics and policy. Funding and research design should include non-academic partners early enough to shape questions, methods, evidence and adoption.

Fanqi Zeng: How can research be transferred more effectively into public knowledge and benefit?

Prepared question; the live discussion ended before a full response.

The wider discussion suggests a practical route: publish accessible summaries alongside academic outputs; involve public, professional and policy audiences in workshops; translate findings into guidance, demonstrations and training; and evaluate whether communication changes understanding or behaviour. Public impact should be designed into the research programme rather than added only after publication.

Professor Wei Jie: What are the key research challenges in applying agentic AI to threat intelligence and cyber-risk assessment?

Prepared contribution; not explored fully during the live session because of time.

Three linked challenges were identified: maintaining reliable assessments as threats and data sources change; embedding transparency, fairness, accountability and other responsible-AI requirements; and sustaining performance over time. Addressing them requires joint work among AI researchers, cybersecurity specialists, social scientists and operational partners.

> **Theme conclusion:** CRANE can create distinctive value by giving ECRs real agency, bringing industry and policymakers into research formation, enabling interdisciplinary teams and translating technical results into public and operational benefit.

Audience questions and follow-up responses

The session closed with audience questions. Trustworthiness metrics and the concept of an audit agent were addressed briefly during the meeting; other questions could not be discussed fully within the available time. The following concise responses consolidate the panel discussion and identify directions for further work.

Audience question 1: What direction should be taken to assess LLM trustworthiness?

Trustworthiness should be decomposed into measurable properties rather than treated as one score. Depending on the use case, evaluation should cover factuality, validity, reliability, calibration, robustness, security, privacy, fairness, explainability and accountability. Tests should include adversarial prompts, distribution shift, domain-specific harms and performance of the full application - including retrieval, tools and human use - rather than the base model alone.

Audience question 2: Is there a role for a dedicated audit agent that reviews the permissions and activities of other agents?

Yes. An audit agent could continuously examine permissions, tool calls, data access, inter-agent communication and behavioural anomalies, while recommending revocation, escalation or investigation. Agents should receive task-specific, time-limited privileges that are periodically recertified. The auditor should not be the root of trust: deterministic policy engines, protected logs and human review must enforce high-impact controls and prevent circular self-assurance.

Audience question 3: Do we need a zero-trust architecture for AI agents?

Yes, especially where agents can access sensitive data or change real systems. Each agent and tool should have a verifiable identity; every interaction should be authenticated and authorised; credentials should be short-lived and scoped; execution should be sandboxed; memory and data should be separated by policy; and consequential transactions should require independent approval. Assurance should combine provenance, threat modelling, red-teaming, continuous telemetry, incident response and safe shutdown.

Audience question 4: How could AI/computer science, applied cognitive psychology and semantic ontologies work together?

This is a promising interdisciplinary direction. Ontologies can represent goals, permissions, provenance, evidence, context and risk in a machine-readable form. Cognitive psychology can examine automation bias, trust calibration, warning fatigue, mental workload and situation awareness. Together they could support context-aware access control, human-centred explanations and experiments that test whether people understand, challenge and safely use agent recommendations.

Audience question 5: Must we accept that comprehensive safeguarding of agentic AI may never be technically achievable?

Perfect safeguarding is unlikely because LLM behaviour is probabilistic, context sensitive and difficult to anticipate in novel situations. This does not justify unmanaged risk. The appropriate objective is bounded autonomy: minimise privileges, contain the blast radius, detect abnormal behaviour, fail safely and retain human authority. High-consequence rules should be enforced by deterministic systems outside the model.

Audience question 6: Is pre-deployment certification fundamentally insufficient for adaptive AI agents? What should runtime assurance look like?

Pre-deployment evaluation remains necessary but is insufficient. Runtime assurance should observe prompts, plans, memory changes, retrieved content, tool calls, data flows and outcomes; enforce policy gates; detect anomaly and drift; require step-up approval for high-risk actions; and support quarantine, rollback and safe fallback. Continuous red-teaming and tamper-resistant evidence should update a living assurance case throughout deployment.

Audience question 7: How is the security community approaching APT and state-sponsored abuse of AI agents and MCP-based systems?

Current work is mapping threats such as indirect prompt injection, goal hijacking, malicious tools or servers, confused-deputy behaviour, credential misuse, memory poisoning, inter-agent impersonation, cross-tool privilege escalation and data exfiltration. Priorities include workload identity, delegated authorisation, signed tool manifests, supply-chain assurance, semantic information-flow tracking, composition testing, runtime detection and shared incident telemetry.

Cross-cutting conclusions

1. **Assure the system, not only the model.** Evaluation must include data, retrieval, memory, tools, permissions, infrastructure, people and operational context.
2. **Treat agent actions as security-relevant transactions.** Each tool call should carry identity, purpose, authorisation, provenance and an auditable result.
3. **Keep deterministic boundaries around probabilistic components.** Models may recommend, but external controls should enforce non-negotiable safety, access and compliance rules.
4. **Design for compromise and recovery.** Systems need containment, anomaly detection, credential revocation, rollback, quarantine and incident reconstruction.

5. **Measure human oversight.** Assurance should test whether people understand, question and can override the system, not merely whether an approval field exists.
6. **Make assurance continuous.** Adaptive agents, changing tools and evolving threats require runtime monitoring and periodic reassessment rather than one-off certification.
7. **Build interdisciplinary and inclusive research structures.** Technical controls will be incomplete without behavioural, legal, ethical, organisational and domain expertise.
8. **Develop a zero-trust reference architecture for agents.** Define identities, scoped permissions, policy decision and enforcement points, secure tool mediation, memory boundaries, approval gates and evidence requirements.
9. **Establish a runtime-assurance research programme.** Create testbeds for monitoring trajectories, detecting drift and manipulation, validating human intervention, and exercising quarantine and rollback.
10. **Investigate supervisory and audit agents safely.** Study their detection value, independence, failure modes and susceptibility to collusion or common-mode compromise, with deterministic enforcement underneath.
11. **Create sector-specific evaluation profiles.** Prioritise healthcare, critical infrastructure and threat intelligence, including domain harms, acceptable error, operator needs and recovery requirements.
12. **Expand human-factors work.** Examine trust calibration, warning design, workload, social engineering and decision-making under realistic time pressure.
13. **Build evidence and provenance into prototypes.** Include tamper-evident logs, versioned policies, data and tool provenance, credential lineage and reconstructable decision records from the outset.
14. **Give ECRs visible leadership roles.** Invite ECR-led sessions, challenge projects, mentoring pairs, writing groups and cross-sector secondments, with ECR representation in priority setting.
15. **Deepen industry and policy participation.** Use partner-defined challenge problems, realistic datasets and controlled test environments, while protecting IP, confidentiality and responsible disclosure.
16. **Translate research for wider audiences.** Produce short practitioner briefings, policy notes, demonstrations and public resources alongside peer-reviewed outputs.

SIG programme and next steps

The panel followed the first quarterly hybrid workshop held at Northumbria University on 1 May 2026. That workshop included a keynote by Ollie Whitehouse, Chief Technology Officer of the UK National Cyber Security Centre, and 24 technical presentations across four sessions. The breadth of subjects demonstrated the value of a forum connecting AI for security with security for AI.

The second hybrid workshop is scheduled for 21 September 2026 at the University of West London's Ealing Campus, hosted by Professor Wei Jie, Dr Alireza Esfahani and colleagues. It will include keynotes

from Dr Meera Sarma and Professor George Loukas, together with short research presentations intended to stimulate exchange and collaboration. Further quarterly workshops are planned for January and May in different UK regions.

The longer-term ambition is an annual AI-Cyber conference combining research presentations with peer-reviewed outputs. The panel's themes provide a strong basis for future calls and working groups, particularly around zero-trust agent architectures, continuous assurance, human oversight, CNI resilience, healthcare safeguards and agentic threat intelligence.

Closing statement

The discussion confirmed that the relationship between AI and cybersecurity is reciprocal. AI offers powerful capabilities for detection, analysis and response, while AI-enabled systems introduce new routes to manipulation, privilege misuse and real-world harm. The appropriate response is neither uncritical adoption nor rejection, but carefully bounded and continuously assured deployment. CRANE's distinctive contribution can be to unite technical, human and governance perspectives while giving ECRs, industry and policymakers meaningful roles in shaping research and impact.

Selected frameworks and source basis

1. NIST AI Risk Management Framework and Generative AI Profile
2. OWASP Top 10 for Agentic Applications 2026
3. MITRE ATLAS - adversarial threats to AI systems
4. Model Context Protocol security best practices
5. Regulation (EU) 2024/1689 - Artificial Intelligence Act
6. UK Automated Vehicles Act 2024