North West Partnership for Security and Trust

The North West Partnership for Security and Trust will be brought to a close later this year as we look at how to expand, evolve and future-proof the way we work with academic partners.

From *September 2026* the NWPST will cease to exist in its current format but its successes will be used to inform a wider collaboration model which will benefit national security. This will include a spectrum of academic research, from blue skies through to applied.

We're proud to say that in the five years since the NWPST was founded – as a partnership between GCHQ and Manchester, Manchester Metropolitan, Salford and Lancaster universities – a wide range of impactful research projects have been funded and facilitated. The partnership has also delivered expert speaker seminars, outreach events, roundtables, knowledge share sessions and events, all designed to facilitate the exchange of knowledge between academia and Government.

Our thanks go to the partners we have collaborated with since 2021. We know that sustaining and scaling strong relationships in the North West is vital to national security and we look forward to future collaboration on a much wider scale.

The purposes and guiding principles of the NWPST were developed jointly through discussions between GCHQ Manchester and the four Universities.

The vision of the NWPST was:

1. To facilitate high-quality and impactful research in and across a wide range of disciplines relevant to UK national security challenges.
2. To build relationships with stakeholders across the North West and throughout the region, including government, industry, and civil society, to identify opportunities to develop new products and services, and enhance the productivity and security of existing provision.
3. To develop and deliver accessible and inclusive skills and apprenticeship provision, to support access to opportunities within the digital and data security sectors for citizens in the North West.
4. To engage with communities, stakeholders and the wider public to inform, explain and educate on issues of trust and security.

Past Activities

Sandpits

+ Show

Algorithmic Bias (2022)

This event ran in June 2022, and subsequently two research awards were granted to cross-institutional, interdisciplinary teams for delivery in 2023/24.

The theme of the Sandpit was "Algorithmic Bias". It is widely recognised that the use of data-driven models for classification and decision making is vulnerable both to the existence of historic prejudicial decision-making reflected in the data, and to the omission of significant communities from the data leading to blind spots in the resulting models. This has the effect both of degrading the accuracy of the classifications and decisions and to the continuation, and in some cases magnification, of injustice. In this sandpit we co-created innovative interdisciplinary projects to explore aspects of algorithmic bias and its mitigation from technical, social, and socio-technical perspectives.

Digital Sustainability (2023)

This event ran in June 2023, and subsequently two research awards were granted to cross-institutional, interdisciplinary teams for delivery in 2023_24.The theme of the Sandpit was "Digital Sustainability". The UK strategy to achieve net zero by 2050 includes a range of proposals from developing green technologies to decarbonising heat and buildings. The paperless office, online meetings, optimising energy consumption and scheduling logistics are all

examples where digitisation can help reduce our carbon footprint. Equally, we know that the cloud and large deep learning models such as ChatGPT consume huge amounts of energy and can have a negative impact on our carbon footprint. The impact of initiatives and policies to reduce energy consumption is also something that needs to be considered to ensure the problems faced by deprived regions are not exacerbated. In this sandpit we co-created innovative interdisciplinary projects to explore the impact of digitisation, and how it can best be used to achieve net zero goals.

Human – AI Teaming (2024)

This event ran in July 2024, and subsequently two research awards were granted to cross-institutional, interdisciplinary teams for delivery in 2024_25. The theme of the Sandpit was 'Human – AI Teaming'. The theme of the Sandpit is "Human-AI Teaming". Human-AI Teaming describes Humans and AI systems working together with meaningful shared control, delivering results which are better than either would be able to produce on their own. Analysts in the National Security Community are faced with ever growing volumes of data and harnessing the data-processing and rapid response capabilities of AI could enable them to focus more on the type of complex and novel analysis which relies on human judgement. Ethical optimised use of Human-AI Teaming will therefore be essential to maintaining both the UK's intelligence advantage and cyber defences, yet there is much still to be determined in order to realise the opportunities of AI while mitigating the risks.

Resilience of Critical National Infrastructure (2025)

This event ran in April 2025, and subsequently two research projects were granted to cross-institutional, interdisciplinary teams for delivery in 2025-2026. The theme of the sandpit was the resilience of critical national infrastructure.

Critical National Infrastructure (CNI) are those critical elements of infrastructure whose loss or compromise could severely impact the delivery of essential services or have significant impact on national security, national defence, or the functioning of the state. It also includes some functions, sites and organisations which are not critical to the maintenance of essential services, but which need protection due to the potential danger to the public (civil nuclear and chemical sites, for example).

Cyber-related

+ Show

Cyber MALCULT

This project analyses the cultures of online groups that seek to do harm through offensive cyber operations and the online dissemination of disinformation and other malicious content. The project's focus is the online cultural profiles and traits of Russian and Chinese Hackers. The project will build understanding of the Beliefs, Ideas, Behaviours, Language and Emotes (BIBLE) of these groups and how they interact with online platforms and technologies. The project will introduce and test a new methodology for studying online security cultures that provides enhanced opportunities to anticipate the actions and intentions of malicious groups. The project draws on expertise and methods in political science, computer science, sociology and behavioural science and provides a detailed empirical analysis of groups involved in malign influence campaigns. It will push the boundaries of understanding of how cultural traits interact with modern computer systems, platforms and networks.

Completed Research 2023

+ Show All

Future of Interlinked computing Phase 1

+ Show

As novel technology moves from science fiction to purchasable solutions, organisations move to adopt novel technology quickly and take advantage of innovative technologies for existing business problems, however this can lead to a cobbling of software solutions, where an organisation no longer has a clear divide between their systems and third parties. This presents many risks ranging from system stability in the case of systems which adapt their behaviour, to security with many of these systems forming an unknown-unknown attack surface. This system-of-

systems involves many organisations with no single locus of control (or potentially of responsibility, accountability or energy management), and will have interactions that cannot be determined by any single participant. Therefore, understanding these complex interacting systems will be key to enabling innovation without compromising system stability, security or environmental impact.

The Security Gene

+ Show

Machine Learning (ML) technology is now at the forefront of societal society, increasingly responsible in aiding human capability to perform knowledge discovery and decision making. Whilst using ML to augment cyber security tools and techniques is gaining considerable academic and commercial traction, less attention has been given to cyber attacks against ML systems itself. Adversarial machine learning are attacks against ML systems that elude classical security techniques, making it possible for attackers to steal/recreate ML model training data, evade classifier detection, produce incorrect ML decision making, and reverse engineer ML model operation.

Completed Research 2024

+ Show All

Identifying Intentional Algorithmic Bias and its Effect on Model Behaviour

+ Show

Unconscious algorithmic bias has received considerable attention in the research community and often causes unintentional disproportionate harm to certain parts of society. However, an important new dimension of bias that has not received such attention is intentional bias, i.e., deliberately introducing bias into a system at the learning stage that can skew the system's behaviour and outcome decisions.

This represents an emerging vulnerability with the potential to significantly degrade performance and can have serious consequences such as discriminatory lending, hiring, and fake news proliferation. This interdisciplinary research project studied intentional bias from a technical and social perspective, including how to identify such bias and its effects on semi-supervised and unsupervised model behaviour in economic and financial scenarios such as AI-based financial hedge funds.

Scoping for Algorithmic Fairness Metrics

+ Show

The use of algorithms to guide decision making is increasingly prevalent—as are concerns about the biases within these algorithms. There is a growing literature that attempts both to mitigate biases and measure fairness (metrics). However, this literature is largely technically-focused, with little attention on how to best communicate these metrics.

This raises two important questions:

1) are technical metrics widely understood, and;

2) are technical metrics for fairness the optimal approach?

This project aims to understand the current landscape: what fairness metrics are currently being used, how well they work, how well they are understood, and what can be improved in terms of measurement and communication.

The research aims to improve the communication of fairness metrics and in turn reduce the misuse of algorithms and thereby promote transparency and trust. Adopting a mixed-method approach, the scoping phase will review relevant literature and elicit user groups' opinions to inform an optional testing phase in which metrics will be evaluated. The findings will be summarised and disseminated at local events to raise awareness and grow interest.

Digital Sustainability: CRESCORDIA

+ Show

An increased dependency upon digital services brings with it a unique series of risks, particularly for modern communities whose essential services, such as healthcare, education, energy use, travel and banking, increasingly rely upon stable digital infrastructures. The digital sustainability of such communities is, then, an increasingly important question when it comes to the security of the economic, social and environmental milieu, particularly when communities face trade-offs between multiple objectives (e.g. between digital security and environmental sustainability). This project seeks, therefore, to ask how modern communities can become digitally sustainable in the face of an increased dependency on digital services. Given the basic premise of sustainability to meet the needs of the present without compromising the ability of future generations to meet their own needs, it is essential to evaluate how communities can learn to build resilience to the pressing demands of both present and future generations into their everyday practices.

Digital Sustainability: Balancing Machine Learning

+ Show

Artificial intelligence (including machine learning) has a significant role in enhancing digital sustainability by improving efficiency and reducing waste. However, it is crucial to ensure that AI applications are developed in a sustainable manner. The rapid growth of the AI market results in substantial emissions. This is mainly due to complex AI models that prioritize accuracy over energy efficiency. Notably, training GPT-3, a large language model, for high accuracy, consumed more electricity than 100 US homes consume in a year. It is worth questioning whether such high accuracy is necessary for every application. Research is needed to understand the energy consumption of creating, training, and operating ML models. This would help us tackling the main drivers of energy consumption related to a models' computational complexity while maintaining acceptable performance and security, we can develop efficient and sustainable digital solutions. We will do so by testing both science and industry cases and develop an end-to-end pipeline for a pilot study from an industry case for an applicable approach to sustainability.

The Security of Artificial Intelligence

+ Show

This project investigates the salient vulnerabilities of AI technologies throughout their life cycle; how attacks, tools, methods, and vectors proliferate in the international system and between societal sectors in a rapidly changing and highly volatile world order and how transferable they are between AI technologies and sectors; and how government, academia and industry can better identify when AI systems have been compromised through a set of common technical and behavioural indicators.

The project draws on political science, strategic studies, computer science and behavioural science competencies to build a network of researchers in the Northwest working on the Security of AI. The project will also help establish and disseminate best practices on the security of AI.

Future of Interlinked Computing Phase 2

+ Show

This project builds on the previous work carried out on The Future of Interlinked Computing. Futurists forecasted the future implications of ubiquitous connectivity, artificial intelligence, mixed reality, low and no-code solutions, and digital ownership, where these challenge UK wellbeing and values. The timeframe considered will be the next 15 years or so, specifically to 2040 and will identify specific ways in which the forecast changes may affect distinct aspects of the 13 CNI sectors and related national interests.

Future of Interlinked Computing Phase 3

+ Show

This project builds on the previous work carried out on The Future of Interlinked Computing Phase 2. Futurists forecasted the future implications of ubiquitous connectivity, artificial intelligence, mixed reality, low and no-code solutions, and digital ownership, where these challenge UK wellbeing and values. The timeframe considered will be the next 15 years or so, specifically to 2040 and will identify specific ways in which the forecast changes may affect distinct aspects of the 13 CNI sectors and related national interests.

Academic outputs:

- (Interlinked Computing in 2040: Safety, Truth, Ownership and Accountability.

- The Human Factor: Addressing Computing Risks for Critical National Infrastructure towards 2040.)

AI Code Repair

+ Show

Standard development libraries focus on delivering efficient AI code, which can often contain security vulnerabilities, necessitating time-consuming fixes for software developers. This project combines Large Language Models (LLMs) with Formal Verification (FV) to ensure standard libraries produce secure and efficient AI code.

Academic Outputs:

- The FormAI Dataset: Generative AI in Software Security through the Lens of Formal Verification
- LLM-Generated Invariants for Bounded Model Checking Without Loop Unrolling
- How secure is AI-generated code: a large-scale comparison of large language models
- SecureFalcon: Are We There Yet in Automated Software Vulnerability Detection With LLMs?
- A New Era in Software Security: Towards Self-Healing Software via Large Language Models and Formal Verification
- UnitTenX: Generating Tests for Legacy Packages with AI Agents Powered by Formal Verification
- VeriExploit: Automatic Bug Reproduction in Smart Contracts via LLMs and Formal Methods

Error Reporting in High Reliability Organisations

+ Show

High-reliability organisations operate with extremely low failure rates in high-hazard environments, such as aviation, nuclear power plants, and healthcare. Error-reporting (e.g., incidents, near misses) in these organisations is critical to organisational learning and continuous improvement. This project will examine the methods that High-Reliability Organisations use to ensure error-reporting and evaluate their effectiveness.

Near-4-Core

+ Show

This project focuses on the challenge of Automated Risk and Policy by exploring the applicability of Natural Language Processing (NLP) techniques to extract qualitative context and quantitative information from a priori risk and policy compliance assessments to support decision making in future operations planning.

This project intersects with Reasoning Over Unstructured Data given the unstructured nature of the risk and policy compliance data forming the corpus used for the NLP. Additionally, we are targeting Improving Cognitive Performance of Operators and Analysts as we will explore the way the resultant large volume of information and intelligence can be visualised at high resolution and accessed at scale in a cutting-edge Decision Theatre.'

Completed Research 2025

+ Show All

Behavioural and Emotional Responses to AI

+ Show

This project explores how Artificial Intelligence (AI) impacts human behaviour and emotions, focusing on cybersecurity professionals, or analogous groups, in high-stakes environments. Unlike AI systems, humans have fluctuating needs, emotions and preferences that may be impacted through interaction with these systems. While AI aids in complex problems, data analysis, and pattern recognition, it can also cause distrust, anxiety, and stress, affecting decision-making. We will review literature, conduct surveys, and map interactions to understand human-AI collaboration.

The goal is to propose strategies and tools for more harmonious and productive Human-AI (HAI) relationships. Key outputs include insight into emotional and behavioural responses of users as they interact with AI systems, an eco-system map showing the relationships between various elements within the AI-human team, and a proof-of-concept

proposal for HarmonyBot, an AI system designed to reduce human's adverse responses and improve user wellbeing. The project ensures inclusivity and fairness, enhancing trust, productivity, and confidence in AI systems.

Perceptions of AI

+ Show

The project conducted a Literature review investigating public perceptions of artificial intelligence (AI), with particular attention to the cultural and demographic factors that influence these opinions. The review also assessed public attitudes toward the application of AI in national security and law enforcement contexts.

Securing AI: Assessing Socio-technical Risks in Artificial Intelligence Models (AISEC phase 2

+ Show

The project aims to generate models of Human-AI collaboration in Reinforcement Learning (RL) which capture the key features of critical tasks relevant to the NWPST, and to design collaboration-aware RL algorithms which behave optimally and robustly in such settings. The work was structured around three Research Questions.

Completed Research 2026

+ Show All

Coordinated Resilience in Critical National Infrastructure: framework for assessing and mapping interdependencies (CRUX)

+ Show

CRUX project aims to improve how the UK protects and manages its most essential systems, known as Critical National Infrastructure (CNIs)—like energy and communications—that we all rely on every day. When one of these systems fails, it can quickly affect others, leading to widespread disruption.

Currently, these systems are often managed separately, making it harder to respond effectively to emergencies. To tackle this, the project will create a new **Coordinated Resilience Framework** that helps organisations work together better. It will also produce a **Cascading Risk Map** to show how problems in one system can impact another, and a **Scenario Simulation Repository** to test how different types of disruptions might play out. Finally, clear **policy recommendations** will be developed to help government and industry respond more effectively to future challenges.

Metrics

+ Show

Resilient CNI can both withstand threats and recover from successful attacks. Given the myriad of threats, from cyber to physical, facing UK CNI understanding the resilience of a sector is essential, however currently there are only limited ways to measure this, and these are inconsistent across different sectors, making it difficult to give an at-scale view of resilience.

The project explores different qualitative and quantitative measures for assessing resilience to cyber, physical, and logistical threats. These will be multidisciplinary, ranging from probability and cybersecurity to human factors and socio-technical, thus providing a comprehensive view of resilience. This will allow analysts to identify weak points in UK CNI, and support decision makers addressing them. The output will be a review of existing resilience measures, identification of a set of metrics designed for CNI, and a proof of concept applying these measures to an emulated 4/5G Terrestrial Network.

Perceptions of Intelligence Agencies among underrepresented youth communities in the North West of England

+ Show

Research shows that trust in the Intelligence Agencies is considered relatively high compared to many other parts of government, but public knowledge of the agencies' remit, is very limited. Research, including the Project Lead's own survey data, has also hinted at a potential generational shift in attitudes. Younger age cohorts are less trusting, less

knowledgeable and much less likely to use traditional media for information about Intelligence Agencies. This project, therefore, seeks to further explore this potential youth dimension by examining five questions:

1. What are the levels of knowledge and trust in Intelligence Agencies amongst young people in North-West England?

2. What are the perceptions of competence of such agencies from young people?

3. What sources of information are used by young people in their assessment of intelligence organisations?

4. Are there differences in the 18-24 age cohort based on other factors (e.g. gender, education level, ethnicity)?

5. What shapes young people's perceptions and attitudes towards Intelligence Agencies beyond any specific knowledge of the agencies?

Expert Speaker Seminars

+ Show

The NWPST sources expert speakers from our partner universities to give seminars to our internal staff on topics of National Security interest. We have held expert speaker seminars on various topics, such as:

- Geopolitics, world events and future crises
- Materials research and innovation in quantum
- AI
- Satellites
- Thriving at Work
- Electricity generation and distribution

PUBLISHED

4 June 2026

REVIEWED

9 October 2024