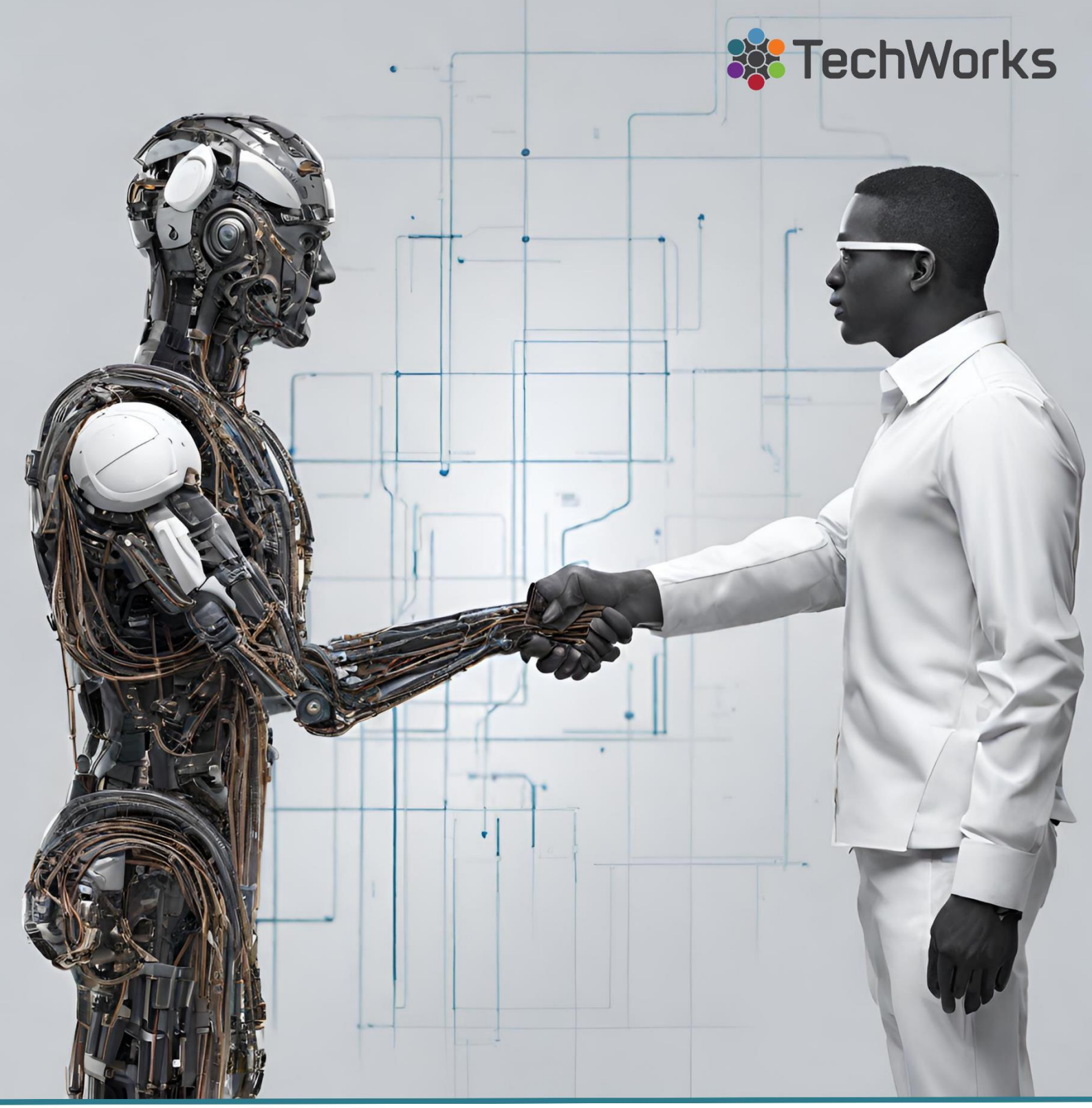
TechWorks



Engineering Trustworthy AI

A TechWorks Communities White Paper

AESIN AUTOMOTIVE ELECTRONICS INNOVATION

IoT Security Foundation

nmi National Microelectronics Institute

TechNES

POWER ELECTRONICSUK

Introduction

Over the summer of 2023, TechWorks ran a number of events with the UK technology sector tackling the subject of Trustworthy AI [1][2]. The events were well attended, attracting representatives from over 60 organisations, from industry bluechips and SME's, academia and trade bodies. Each party has a self-declared interest in the opportunities and challenges that AI represents.

TechWorks' is the UKs Deep Tech Hub which hosts many technologists, engineers, entrepreneurs, researchers and business visionaries hence the events had an innovation and engineering focus. Our ambition is to identify opportunities to collaborate on the technical building blocks that will deliver on the vision of trustworthy AI systems and hence "Engineering Trustworthy AI" was the premise under which these meetings were convened.

The discussions were both deep and wide ranging, covering many different perspectives. In this paper, we summarise the key findings, and circulate them for wider awareness and discussion.

Common Trustworthy AI Themes

Trustworthy and responsible AI are garnering much interest - within governments globally, within intergovernmental organisations, and within corporate boardrooms. The result has been multiple initiatives, each with a slightly different set of objectives and requirements, but all of which are based on a similar set of principles.

For this paper, we reference the NIST [3] principles, which are useful and broadly representative:

- Validity and Reliability
- Safety
- Security and Resiliency
- Accountability and Transparency
- Explainability and Interpretability
- Privacy
- Fairness with Mitigation of Harmful Bias

At a conceptual level, these cover the most critical issues that an AI system must address to be considered trustworthy. At the convened Techworks workshops we sought specific, concrete, practical advances that can 'demonstrably' move the dial on each of these dimensions.

Hence, our starting questions:

1. What can we learn from existing (non-AI) trustworthy systems and what learnings might we apply to emerging AI systems?
2. What is specifically challenging from an AI systems perspective and what new initiatives and/or innovations are necessary requirements?

Informed by recent work from our members, we have been keen to explore the potential of new forms of Bills of Materials (BoMs) and their utility in the development of trustworthy systems. Software Bill of Materials (SBoMs) have gained prominence in recent times - boosted by the US Executive Order 14028 from May 12, 2021 and their significance in relation to 'improving the nation's cybersecurity'. We thus conceive, and consider the notion of an AI Bill of Materials (AIBOM) as a foundational concept in this paper.

Key Discussion Points

The wide-ranging discussions that took place across the workshops revealed a great many facets to this frontier technology and here we attempt to distil them into a simpler form. We have grouped the discussion into related themes. In this paper we do not seek to priority order the themes as, by the nature of a complex system, each component is significant and inter-related:

- Scope and definition of AI
- Research, Skills & Training
- Explainability and Transparency
- Bias
- Infrastructure
- Training Data
- Complexity & Dependencies
- Versioning
- Best Practice
- Testing
- Liability
- Provenance
- Quality and Quantity of Data
- Context
- Subjectivity
- Failing Safe
- Trustworthy Hardware
- Trustworthy Software

As noted previously, our primary interest has been to ask what can the TechWorks community contribute in practice? We consider four possible conclusions:

a) The problem space is intractable in the near term and represents a research grade problem;

b) The challenge is being handled effectively by other initiatives;

c) Solutions are expected to be outside the Techworks' collective competence to be addressed readily; and

d) It is an actionable problem where the Techworks community can contribute.

The themes of the discussions at the workshops are summarised in the following list together with recommendations.

Scope	**What is AI?** How do we define an AI system? What is the difference between AI, statistical methods and other mathematical algorithms? Do we care? The issue becomes of vital importance, when applying legislation; you need to know if your system is subject to it. **Conclusion:** Outside Techworks immediate competence. **Recommendation:** Monitor progress and communicate to membership as appropriate.
Research, Skills & Training	**Does UK industry have access to necessary capabilities?** Indisputably a vital issue for research and economic development. At the research level, there are nascent and existing programmes, relating to CDTs and AI hubs that look promising. Yet, for UK industry, it is not so clear what is happening or difficult to keep abreast of. **Recommendation:** monitor evolving landscape, liaise across TechWorks communities and stakeholders to gain awareness of initiatives and further surface gaps in what is needed. Disseminate/communicate in simple form as necessary.
Explainability (Transparency)	**Can the outputs of an AI system be explained? Is there transparency in the decision process?** This is one of the key concerns of trustworthy AI and appears in most commercial and (inter-)governmental work on trustworthy AI. However, for large scale LLM and neural network systems, it is unclear how this is achieved practically at an operational level. It is still a complex research grade problem. **Conclusion:** Outside Techworks immediate competence. **Recommendation:** Monitor progress and communicate to membership. *Note: our curiosity as to an AIBOM has the potential to digitally record information relating to explanations of specific inferences, however it does not fundamentally solve the problem of generating those explanations.*
Bias	**How do we ensure fairness and prevent bias?** No one disagrees with the aspiration however it is a complex problem. Doing this in practice is difficult and considered outside of the technical challenges of designing and delivering the system. There is a related and fundamental challenge of how to gauge or measure bias. This issue also hits on the **subjectivity** (i.e. who is measuring the bias) and **contextual** (i.e. bias for which application). **Conclusion:** Outside Techworks immediate competence.

	Recommendation: Monitor progress and communicate to members. Ensure the measurement requirements are captured for any practical work. *Note: the AIBOM approach has the potential to digitally record the results of tests for fairness and bias that might occur as part of an assurance process.*
Infrastructure	**Does the UK have the infrastructure to train large AI models?** Training state-of-the-art AI systems, such as LLMs, requires significant (i.e. massive) amounts of compute and is hugely expensive. However, without the ability to train models there is little that can be done to assure their trustworthiness, and realisable innovation cannot occur. Note: There have been some recent developments: examples include a £500 million UK government sponsored initiative [4] and a £2.5 billion investment by Microsoft [5]. **Recommendation:** Monitor progress and communicate to membership. Lobby to ensure that any practical trustworthy AI initiatives are measurably embedded within the proposed infrastructure.
Training Data Consent	**Is there a notion of consent to train?** What rights do data owners have when AI systems have been trained on their data? This is a complex problem that will likely be contested in the courts across multiple jurisdictions/legislations. In addition to court cases that are already in progress, a number of collaborative initiatives have been created to help address these issues. **Recommendation:** Monitor and communicate. *Note: an AIBOM approach has the potential to digitally record consent in the AI system descriptors.*
Complexity (Dependencies)	**Compared to traditional/contemporary systems, how complex are AI systems?** AI systems are considered to be significantly more complex than conventional software. This is because they inherit the complexity of normal software management and in addition, their behaviour can be determined by literally billions of configuration parameters at run time. These configuration parameters are in turn derived from a complex algorithm (different software to the operational software), trained from billions of bits of data, under process in term configured by training parameters. The resulting system 'of many moving parts', still depends on the integrity and security of the training system (hardware, operating system and software) and the operational system, both of which can be highly distributed. **Conclusion:** These are considered as statements of fact. There is no

	direct action required, other than to ensure true understanding of this complexity is represented in any initiative undertaken. *Note: an AIBOM approach does not solve this problem, but offers a precise language to describe the problem.*
Versioning	**How do you know if an AI system's behaviour will be fundamentally changed?** What version of an AI system are you working with? If you are an AI system provider, trust in your system will depend on your users ability to determine the version they are using. If you are the user of an AI system, you need to be able to determine and record the version that you are using. This is important for matters such as liability, transparency, traceability, robustness, etc. **Recommendation:** embedding the concept of versioning directly into the proposed AIBOM activity will be essential to support trustworthiness.
Best Practice	**How can we encourage best practice amongst AI system developers and AI users?** This issue is garnering a lot of attention from both industry and government. At the general level of best practice and process, there would appear little that can be usefully added, until the dust settles. However, within the specific vertical of electronic systems, Techworks is uniquely positioned to guide and support engineering in an increasingly AI dominated landscape. It can do this by providing guidance on best practices in AI that is oriented toward those already embedded in the electronic systems space. Through this, Techworks can both help electronic systems engineers both thrive in the rapidly growing field of AI, and potentially give back to that field by building skill sets that can address talent shortages in this rapidly growing field. **Recommendation:** Techworks should continue to produce and refine guidance for best practices in AI oriented at electronic systems engineers.
Testing	**How do we confidently test an AI system?** Testing AI systems is complex. What system are you testing? What are you testing the system against? What part of the AI system are you testing? Who is doing the testing? Which version are you testing against? What is the application to which the AI system is being put? These questions highlight the issues of complexity, versioning, context and subjectivity as discussed in this article.

	There are likely lessons we can draw from industry sectors here such as semiconductors, embedded systems, IoT and automotive. Digital signing technologies allow us to independently sign the test results of component parts and infer the trustworthiness of the entire system from its parts. **Recommendation:** testing should be explicitly added as a use case to be addressed in the AIBOM activities, to be investigated.
Liability	**Who is liable if/when damage is done by an AI system?** As with all things AI, this becomes a very complex question. Not only does an AI system have many moving parts, each moving part can have a different stakeholder or legally liable owner. Determining liability can be a slow/complex process of: identifying the moving parts, identifying the component owners, identifying the explicit legal relationships that exist between these owners, identifying the implicit legal relationships that exist between these owners, identifying issues of consent and disclosure plus the normal legal issue of identifying damage. **Recommendation:** Monitor and raise awareness of rulings/initiatives in this area. Examine the utility of AIBOM to document component dependencies and the potential for legal relations between parties.
Provenance	**Where did your AI system come from?** The AI provenance question is partly covered by the versioning and complexity issues raised above. Determining provenance is also essential in addressing the challenges of transparency and liability. **Recommendation:** ensure any activities undertaken relating to versioning and dependencies practically address common provenance use cases.
Quality and Quantity of Data	**How do we ensure supply of good quality and requisite quantity of data for the UK industry?** AI performance is heavily dependent on the quality and quantity of data that is used for training. Hence this is a wide ranging issue which is international in nature and spans industry sectors. It also touches on the issues of consent and legal relations between component owners. **Conclusion:** There is little/nothing that Techworks can immediately add to address this problem. However, testing strategies that can give digital certificates to data (to be used for training) will provide a level of transparency and quality control. **Recommendation:** Monitor and communicate. Ensuring the use case of labelling training and test data digitally is part of the use cases addressed in the proposed AIBOM activities.

Context	**What is the system being used for?** Trustworthiness of AI systems is heavily dependent on the context of use: Example 1: A general purpose algorithm to detect pedestrians has an entirely different operational “trustworthy” tolerance if used to count pedestrians crossing a road in contrast to being used within an autonomous vehicle’s control system. Example 2: An AI system to compute APGAR scores in new born babies is entirely dependent on the assessment population [6], and demonstrably so, as those scores have been proven to have implicit racial bias. **Conclusion:** Any trusted engineering innovation to measure trustworthiness of AI systems must be able to express the context of use within the assessment process **Recommendation:** research current best practice for digitally asserting application or context of use. Ensure that any AIBOM activities provide a method for naming and validating applications and content of use.
Subjectivity	**Are assertions of trustworthiness inherently subjective?** A claim made by the originator of an AI system is different to a claim made by an independent test of a system. The level of trust a user has in an AI system will also vary depending on their confidence in the entity that created or tested the AI system. Even when we consider basic system (non-AI) trustworthiness it is clear that subjectivity is an integral part of the evaluation. Take a telecoms system example: Does “China Mobile” trust Huawei telecommunication software? Does Vodafone trust Cisco telecommunication software? But does Vodafone trust Huawei equipment and does China mobile trust Cisco equipment? A robust mechanism for asserting AI trustworthiness must embody the methods that acknowledge this inherent subjectivity. **Recommendation:** any reliable method to describe an AI system must explicitly model the asserting parties, the consuming parties and the trust relations between the two.
Failing Safe	**How can we ensure our AI system fails in a controlled/defined manner?** When we apply an AI system to a specific application, the consequences of failure must be an integral part of its trustworthiness. Fail safe design is an important part of best practices. There are two problems worth separating:

	• Making AI fail safe. This is a challenging, ongoing research problem and considered out of the scope. • Making systems that use AI fail safe. This is still a hard problem, yet more oriented toward general good engineering practices. Best practice in fail-safe design builds on the 'context' issues already raised, but must further manage the failure-pay-off matrix in the context of use. **Recommendation:** Fail-safe design to be embedded into best practice. Attempt to build in fail-safe measurements into any "context of use" mechanisms created.
Trustworthy Hardware	**Trustworthiness is dependent on the hardware we train and run the AI systems on.** Trustworthy hardware is a complex problem in its own right. The security and resilience of the AI system is dependent on the trustworthy qualities of the hardware platform. There is substantive prior art in this area, particularly across the full spectrum of Techwoks competence. **Recommendation:** Trusted hardware best practice and evaluation should be considered as a component element of full AI descriptors to be developed.
Trustworthy Software	**Trustworthiness is dependent on the software we train and run the AI systems on.** Trustworthy software is equally a complex problem in its own right. As per the hardware platform, the 'AI system' inherits the trustworthy qualities of the dependent software. Software Bill of Materials (SBOM) [7] represents current best practice for asserting software dependencies and forms a basis of a user determining the trustworthiness of the assembled system. **Recommendation:** SBOM should be considered as a base component of any developed system for asserting or evaluating the trustworthiness of an AI system.

A StrawMan Proposal

There are significant benefits to deploying nascent AI systems across social and economic sectors and there is a necessary global ambition to increase the trustworthiness of those systems. If technology providers cannot provide trust assurances that AI is safe - to an acceptable and demonstrable level - in use, realisation of those benefits will be impeded. This raises two important questions:

- How do we begin to define and measure trustworthiness?
- How can we delineate the AI system we are measuring?

TAIBOM (Trusted AI Bill of Materials) is a proposed Techworks initiative to look at specific and purposeful engineering principles that can help address these issues. Ideally, such a system should provide:

- A method for defining the immutable properties of a complete AI system; defining a stable AI system, its dependencies and their provenance (where relevant); and
- A method for making and evaluating objective AND subjective claims about the trustworthy attributes of a stable AI system and its constituent parts.

This will be investigative work which will leverage the combined expertise of the TechWorks community. It will highlight the state-of-the-art and elicit the key requirements, with the intention of making solid proposals.

In particular, there are published works from within the TechWorks communities including the IoT Security Assurance Framework[8], SBOM whitepaper[9], and Supply Chain[10] considerations that can directly inform this work, as well as work on Secure by Design[11] principles, and most recently, D3 (Distributed Device Descriptors)[12], which can provide a foundation for digitally describing complex systems from multiple parties.

Conclusion

Evaluating the trustworthiness of an AI system and making assertions about it is a complex challenge - AI systems have many dynamic moving parts which are subjective and context dependent. However, the nature of the problem, and its possible solutions, is an area in which the combined Techworks community has significant breadth and depth of knowledge and experience. We therefore conclude that trustworthy AI is ***an actionable problem where the Techworks community can contribute*** [see section Key Discussion Points above].

Based on our preliminary exploration and recent member events, we propose the follow-on activities for Techworks members:

- **Innovation Cross Working Group:** create a TechWorks cross working group (xWG) to further develop the ‘Engineering Trustworthy AI’ systems concept.
- **Priority Themes:** To commence collaborative working, we will look to
    - Create a Developer Best Practice Guide &;
    - Develop a Trustworthy AI Bill of Materials to support the assertion, demonstration and evaluation of claims of trustworthiness;
- **AI stakeholder map:** We also see a need to simplify and communicate to members the contemporary and dynamic landscape that affects their business. We propose the creation of a practical reference document that will help communicate the different

parties that are seen to exist in relation to an AI system and the position/relationships that exist;

- **Skills:** investigate the opportunity to work with, build on, and extend the success of the UK Electronic Skills Foundation to address the UK's industrial AI challenges; and
- **Knowledge exchange:** construct fit-for-purpose mechanisms which effectively summarise and share information across the Techworks membership.

References

[1] https://www.techworks.org.uk/event/engineering-trustworthy-ai

[2] https://www.techworks.org.uk/event/engineering-trustworthy-ai-workshop

[3] https://www.nist.gov/trustworthy-and-responsible-ai

[4] https://www.gov.uk/government/news/science-innovation-and-technology-backed-in-chancellors-2023-autumn-statement

[5] https://www.gov.uk/government/news/boost-for-uk-ai-as-microsoft-unveils-25-billion-investment

[6] https://www.bmj.com/content/382/bmj.p1620

[7] https://www.cisa.gov/sbom

[8] IoT Security Assurance Framework

[9] Software Bills of Materials for IoT and OT Devices

[10] Securing the Internet of Things Supply Chain

[11] Secure Design Best Practice Guides

[12] https://specs.manysecured.net/d3

About TechWorks

Technology innovation has changed and is changing our world at an enormous rate creating new opportunities across all areas of industry and commerce. With this whirlwind of change comes huge political and societal changes, disrupting the established norms in all walks of life.

TechWorks is a new type of industry association at the core of the UK deep tech community with an ambition to harness our fantastic engineering and innovation to develop the UK's position as a global technology super-power.

TechWorks operates beyond established silos by:

- Creating dynamic, connected technical and business communities to empower innovation and collaboration, supporting business growth and investment.
- Identifying the critical common challenges and leading responses to tackle them.
- Developing the UK tech ecosystem and partnerships across industry, academia and government to ensure the UK is amongst the best location globally to start, build and scale a Deep Tech organisation.

TechWorks

The UK Deep Tech community

www.techworks.org.uk

AESIN
AUTOMOTIVE ELECTRONICS INNOVATION

Dedicated to accelerating Electronic Systems Innovation in the Automotive environment

www.aesin.org.uk

IoT Security Foundation

The Home of IoT Security

www.iotsecurityfoundation.org

nmi
National Microelectronics Institute

Champion for the UK Electronics Manufacturing Industry

www.nmi.org.uk

TechNES

Technology Network for Embedded Systems

www.technes.org.uk

POWER ELECTRONICSUK

Empowering the UK Power Electronics Industry

www.power-electronics.org.uk