Security Conference

The Challenge of Standards for Securing AI - the Work of ISG SAI

Presented by: Scott W CADZOW

19th October 2023

Agenda and outline

- Overview of ETSI’s AI work and what has driven it
  - A couple of important announcements too
- What is the problem with AI? How is ETSI addressing this on our behalf?
  1. Rationalising the security problem of AI
  2. Mitigating the AI security problem
  3. Test and assurance of AI security
- What we’ve done and where we’re going
- Verticals and horizontals

News regarding ETSI and AI

- After 4 years in a pre-standards role as an Industry Specification Group ISG SAI is closing later in 2023
- Phoenix like, we will become ETSI TC SAI in December 2023
- Already on the ETSI portal:
  - https://portal.etsi.org/tb.aspx?tbid=913&SubTB=913#/
- First and second meetings already open for registration:
  - https://portal.etsi.org/tb.aspx?tbid=913&SubTB=913#/5069-meetings
- Mailing lists need to be re-subscribed:
  - Via the mailing lists pane on the portal

ETSI
The Standards People

ETSI Artificial Intelligence (AI) Conference

Status, Implementation and Way Forward of AI Standardization

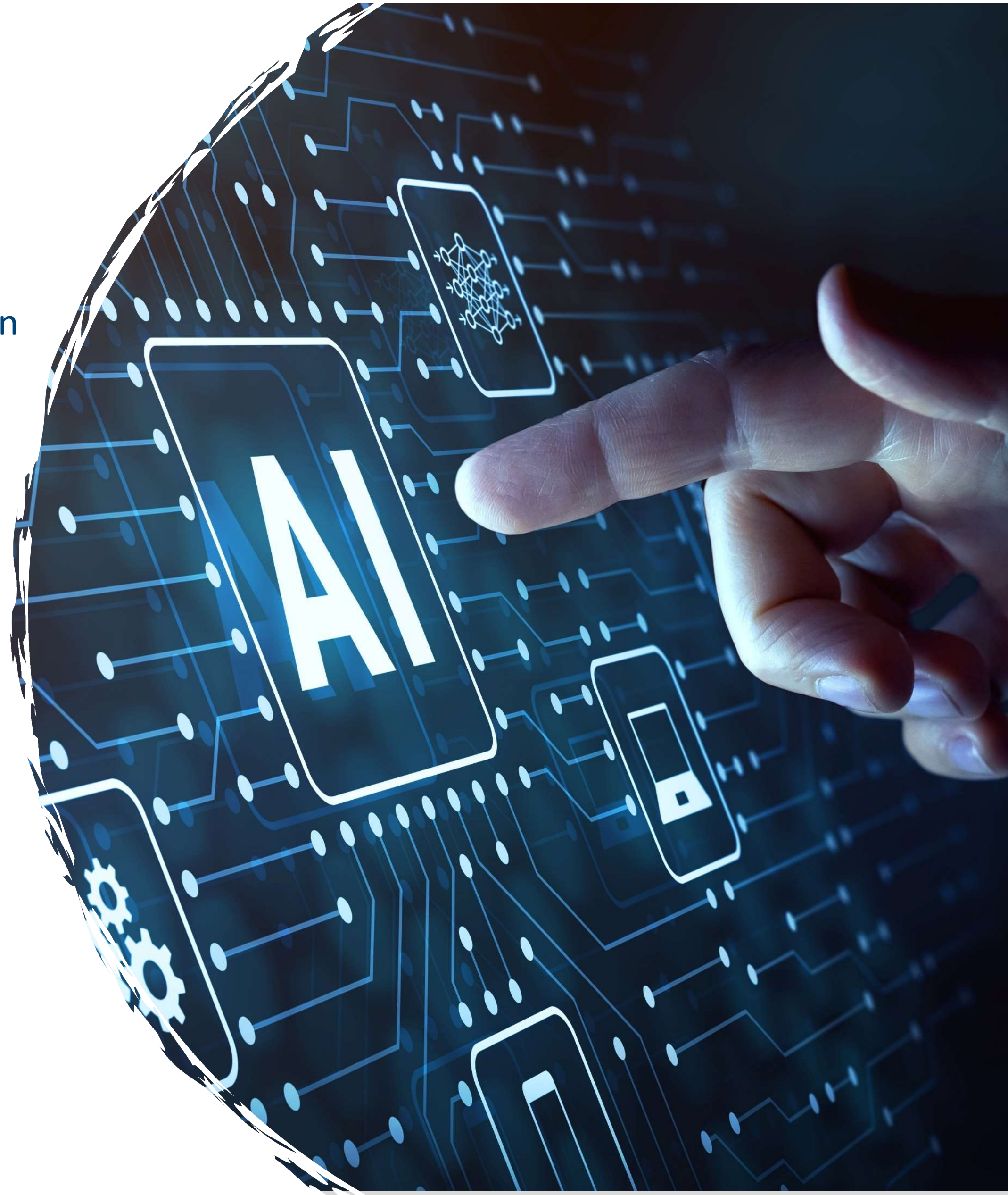
The ETSI AI Conference will focus on AI/ML from the Information and Communications Technology (ICT) perspective.

5-7 February 2024

ETSI, **Sophia Antipolis,** France (F2F)

Information and Registration:
www.etsi.org/etsiaiconference

☞ **Call for Demos and Posters Open!**
Submit your proposals **by 31 October 2023**



What did an AI suggest? (Edge/Bing AI engine)

Securing AI systems presents several challenges, particularly in the domain of cybersecurity. Here are some of the key challenges:

1. **Lack of Transparency and Explainability**: AI systems often operate as a “black box,” making it difficult to understand how they make decisions.
2. **Overreliance on AI**: Overdependence on AI systems can lead to vulnerabilities if these systems are compromised.
3. **Bias and Discrimination**: AI systems can unintentionally perpetuate biases present in their training data, leading to discriminatory outcomes.
4. **Vulnerability to Attacks**: Cybercriminals can manipulate AI systems, leading to threats that can disrupt corporate systems and inflict significant damage.
5. **Lack of Human Oversight**: Without proper human oversight, AI systems can make decisions that have unintended consequences.
6. **High Cost**: Implementing and maintaining secure AI systems can be expensive.
7. **Privacy Concerns**: AI systems often require large amounts of data, which can lead to privacy concerns.

Moreover, the speed of AI developments, parsing the components of what to regulate, and determining who has the authority are also significant challenges for regulating artificial intelligence. These challenges highlight the need for robust standards for securing AI.

The problem?

The unsolved problem set of philosophy: *What exactly is intelligence? How could an external observer prove that an agent is intelligent?*

Superficially there is enough content on the internet to allow a reasonable search to find enough data to construct a good facsimile of what a fairly intelligent human would be able to construct in answering any query

- But it'll do it in less time (seconds compared to hours or days)
- It'll cite its sources if asked too (gives a hint of academic rigour)
- It'll mostly be good enough to pass muster

If you choose to restrict the data used to allow the AI to construct its answer it'll be biased in favour of its sources

- If you only have one book as your world view then your world view will not accept anything not in the book
- Well chosen data acts as a mirror and amplifier of viewer/reader prejudice and can react to feedback to further amplify its effect

The perception problem?

Data manipulation, behavioural assessment, behavioural and opinion manipulation

- Is what you see real?
- Is what you read true?
- Is what you feel being manipulated?

At a leadership level (Government) perceptions are critical and citizens have to be protected

- Elections have to be honest and any threat of manipulation has to be addressed
- Legislation will be enacted to make providers/developers liable for any damage caused by malicious AI

Mitigating the AI security problem

Knowledge, proof, verification are at the heart of making AI secure

- Knowing what the data should be and where it comes from -- Supply chain integrity and provenance
- Knowing what the algorithm should be doing and is doing it

The core conventions of the CIA paradigm still apply → Reinforced by wider adoption of the zero-trust model answering questions of What — Why — When — How — Where — Who

- Authenticate and verify authorisation → Least persistence and least privilege models
- Verifying integrity is not going to be as straightforward as the data is always changing

Notably AI will need to police itself - without interfering with the core functions

- This is not new - we already use software to protect software but it’s a new level of protection we’re seeking

A summary of ETSI’s AI activity

The most recent snapshot of all of ETSI’s AI activity is in our white paper:

https://www.etsi.org/images/files/ETSIWhitePapers/ETSI-WP52-ETSI-activities-in-the-field-of-AI-B.pdf

Lots of history but a few key-points:

- October 2016: White paper describing the Generic Autonomic Networking Architecture (GANA)
- October 2017: White Paper from that set the foundation for ISG ENI
- 2019: eHealth use case publication acknowledges the role of AI as a health professional
- Dedicated technical committees: ISG ZSM, ISG ENI, TC MTS AI Testing, ISG SAI, TC CYBER …

Some of the problems we're trying to solve or resolve

… to define what would be considered an AI threat and how it might differ from threats to traditional systems.

There was no **common** understanding of what constitutes an attack on AI and how it might be created, hosted and propagated when ETSI started in ISG SAI.

- Lots of scare stories inspired by novels, TV, film (the Skynet or HAL-9000 scenario)

As a standards body we need to provide a narrative that should be readily accessible by both experts and less informed audiences across the multiple industries and stakeholders in AI to allow for understanding and debate around the problem.

- Debunking the fictional narratives and promoting a rational discussion

It is essential to address AI in as many forms as it can take: as a system in its own right (rare); as a component of a system; as an adversarial attacker; and, as a system defender

The guidelines for AI governance – OECD

The OECD AI guidelines are a broad stroke attempt to guide the market:

- *AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being.*
- *AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society.*
- *There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them.*
- *AI systems must function in a robust, secure and safe way throughout their life cycles and potential risks should be continually assessed and managed.*
- *Organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles.*

Forms of intelligence (Howard Gardner)

- **Linguistic** - sensitivity to spoken and written language, the ability to learn languages, and the capacity to use language to accomplish certain goals.
  - This is the root of LLMs and includes many of the attributes of generative AI
- **Logical-mathematical** - the capacity to analyze problems logically, carry out mathematical operations, and investigate issues scientifically.
  - Useful in many domains but seeing lots of application in drug development and in diagnostic medicine
- **Musical** - involves skill in the performance, composition, and appreciation of musical patterns.
  - Autotune is the tip of this iceberg and most “studio” environments have the ability to generate auto-bands, auto-orchestras to augment a solo player’s performance
- **Bodily-kinesthetic** - entails the potential of using one's whole body or parts of the body to solve problems.
- **Spatial** - involves the potential to recognize and use the patterns of wide space and more confined areas.
- **Interpersonal** - is concerned with the capacity to understand the intentions, motivations and desires of other people.
  - The worry centre for many people - used to shape and direct interpersonal relationships
- **Intrapersonal** - entails the capacity to understand oneself, to appreciate one's feelings, fears and motivations.
  - The other worry centre - the machine as personal embodiment of the self-diagnostician

Coming to a conclusion as to why we need work

Intelligence is complex

Artificial intelligence is currently man-made … future AI will diverge from man-made constructs (machine "learns" from its failings to design the next machine and in a few generations the human generated root is lost)

- Standards may allow that divergence to be controlled or understood

Humans have traditionally/historically done very badly with "different"

The existence of AI impacts the whole of society. To manage AI it is not just AI that needs to be managed: how human society interacts with "different" intelligence has to be managed. AI and Human Intelligence both need to consider how they work together.

Transversal requirements

Mainly addressed in 3 ETSI groups (and one oversight group)

- ~~ISG~~ TC SAI -- Security across all forms of AI, initially focussed on ML
- TC CYBER -- Integrating advice from SAI into the risk analysis method (TS 102 165-1) and the frameworks for countermeasures (TS 102 165-2)
- TC MTS -- Addressing the role of AI and testing

OCG AI coordinates across all of ETSI for policy and other matters

~~ISG~~TC SAI’s remit addresses terminology, threat and countermeasure, supply chains, privacy, ethics, liability and everything else. Mainly giving guidance -- but now moving into the more technical domain where normative standards will be applied.

Treatment of risk from AI

The matter of calculating risk in AI systems has been addressed in guidance and now in more depth.

- Guidance is given in ISA SAI-001 - the Ontology
- Detail of how AI impacts risk is being embedded into the TVRA method (TS 102 165-1).

Consistency in understanding of risk and identifying countermeasures enables effective vertical sectorial treatment of that risk.

Concern was noted that some language equates risk and impact without due consideration of the likelihood of an event. This has been addressed by slightly modifying the definition of impact and the metrics assigned, without alteration of the metrics and calculation of likelihood. However, the role of motivation, and how AI impacts it, is likely to change the effective likelihood. This is one area under active study to update TS 102 165-1 and then feed the recommendation to all TBs in ETSI. This treatment and discussion of the role of AI in risk assessment will be shared with other SDOs.

The vertical domain

For AI this tends to be in closed environments. ETSI has a large number of projects and TBs ensuring that the data to support machine processing is available in a timely, consistent and semantically described manner.

AI machinery can be applied to sector specific topics such as network resource optimisation (e.g. in ENI for experiential optimisation, or ZSM for autonomous network management), forward event planning and anticipation (e.g. in ITS for traffic flow management)

The role of SAREF as a semantic tool in describing events is key to many sectors (SAREF enables ontologies to allow machine processing)

In summary

- AI is not a problem per-se - it is symptomatic of a desire to use new technology to stretch the horizon of what can be achieved
  - AI can be used for good - x-ray images can be scanned to identify cancers
  - AI can be used for malicious purposes - photos can be manipulated to present a false and damaging image
- The core technologies are almost identical - the tool (AI) is not the problem, rather it is the intent and purpose to which the tool (AI) is applied
- Human behaviour is very often selfish and destructive and has used tools historically to amplify such selfish and destructive traits

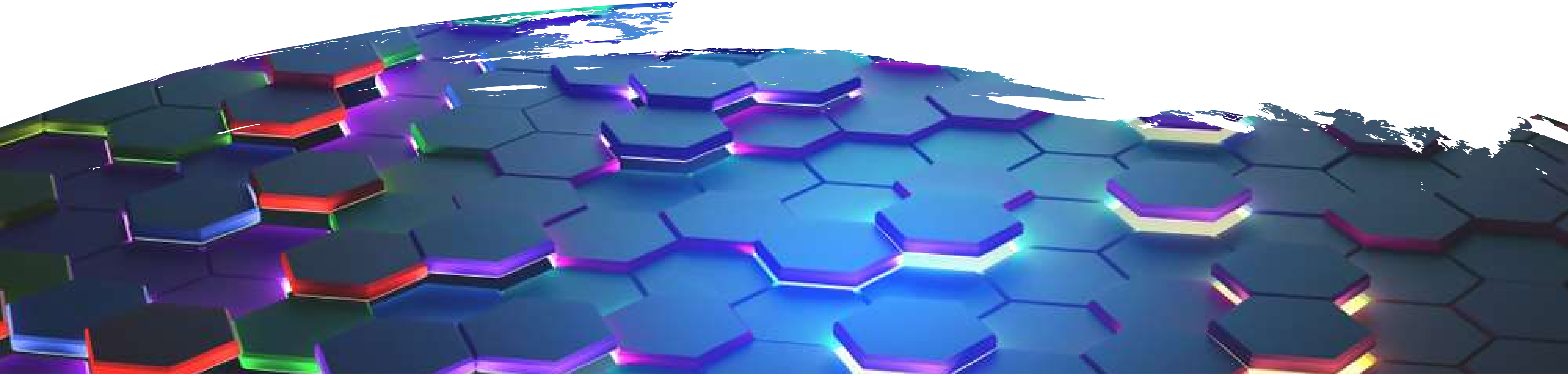
Thank you for your attention

Follow us on:

Any further questions?

Contact me:

scott at cadzow dot consulting



Scope of the AI act

Applies to the EU market

Applicability of the AI act

Sets the framework for global action