ETSI ISG Securing Artificial Intelligence

Presented by: **SAI**

For: **ETSI public portal**

12.2021

ETSI Securing Artificial Intelligence (SAI) Introduction

•ETSI ISG SAI is the first technology standardization group focusing on securing AI. It was officially formed in September 2019; 4 plenary meetings per year and fortnightly interim calls. In 2021 its ISG mandate was renewed for a further two-year term.

- The primary responsibility of our Industry Specification Group on Securing Artificial Intelligence (ISG SAI) is to develop technical specifications that mitigate against threats arising from the deployment of AI – and threats to AI systems – from both other AIs and from conventional sources.

- As a pre-standardisation activity, the ISG SAI is intended to frame the security concerns arising from AI and to build the foundation of a longer-term response to the threats to AI in sponsoring the future development of normative technical specifications.

- ISG SAI aims to develop technical knowledge that acts as a baseline in ensuring that AI systems are secure. Stakeholders impacted by the activity of the group include end users, manufacturers, operators and governments.

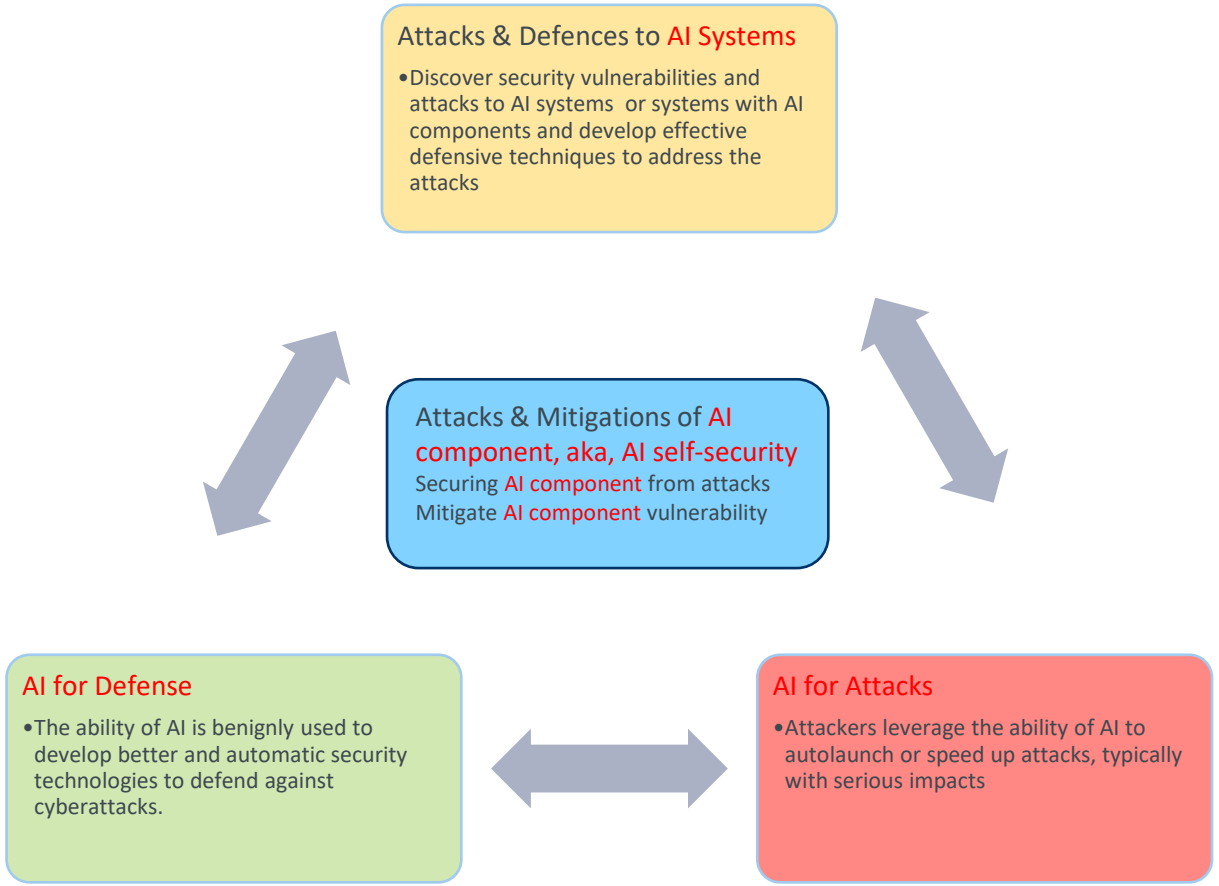
- With the Upcoming EU AI Act the role of ISG SAI may change in the future to a sub-working group of ETSI TC Cyber if mandated standards work is required as this will enable SAI to work on them.

ETSI ISG SAI Scope

The rationale for ISG SAI is that autonomous mechanical and computing entities may make decisions that act against the relying parties either by design or as a result of malicious intent. The conventional cycle of risk analysis and countermeasure deployment represented by the Identify-Protect-Detect-Respond cycle needs to be re-assessed when an autonomous machine is involved.

The intent of the ISG SAI is to address 3 aspects of AI in the standards domain:

1. Securing AI from attack e.g. where AI is a component in the system that needs defending.
2. Mitigating against AI e.g. where AI is the 'problem' (or used to improve and enhance other more conventional attack vectors)
3. Using AI to enhance security measures against attack from other things e.g. AI is part of the 'solution' (or used to improve and enhance more conventional countermeasures).



ETSI ISG SAI Status Quo

Chair:	Alex Leadbeater (BT)
Vice-Chair:	Dr. Tieyan Li (Huawei)
Vice-Chair:	Dr. George Sharkov (SBS)
Secretary:	Alexander Cadzow (C3L)
Technical Officer:	Sonia Compans (ETSI)

Five Active Work Items with Rapporteurs:

- Security Testing of AI: Martin Schneider, Fraunhofer FOKUS
- Explicability and Transparency of AI Processing: Scott Cadzow, Cadzow Communication
- Privacy aspects of AI/ML Systems: Alec Brusilovsky, Inter Digital
- Artificial Intelligence Computing Platform Security Framework: Yu Zhang, Huawei
- Traceability of AI Models: Katarzyna Kapusta, Thales

Published Work Items

GR SAI 004 Securing AI Problem Statement (Group Report)

Scope

This work item aims to describe some of the main challenges of securing AI-based systems and solutions. including challenges relating to data, algorithms and models in both training and implementation environments. The focus will be on challenges which are specific to AI-based systems, including poisoning and evasion.

Motivation

•Practical AI systems have been implemented and enabled by: (1) **Evolution of advanced AI techniques** including neural networks, deep learning (2) **Availability of significant data sets** to enable robust training (3) **Advances in high performance computing** enabling highly performing devices and the availability of hyperscale performance through cloud services (4) These advances primarily relate to **machine learning**, but what about other areas like **reasoning**.

• These new techniques and capabilities, together with the availability of data and compute resources, mean that AI systems will only become more prevalent. However, AI systems have some different challenges which are different from traditional SW/HW systems.

GR SAI 001 AI Threat Ontology (Group Report)

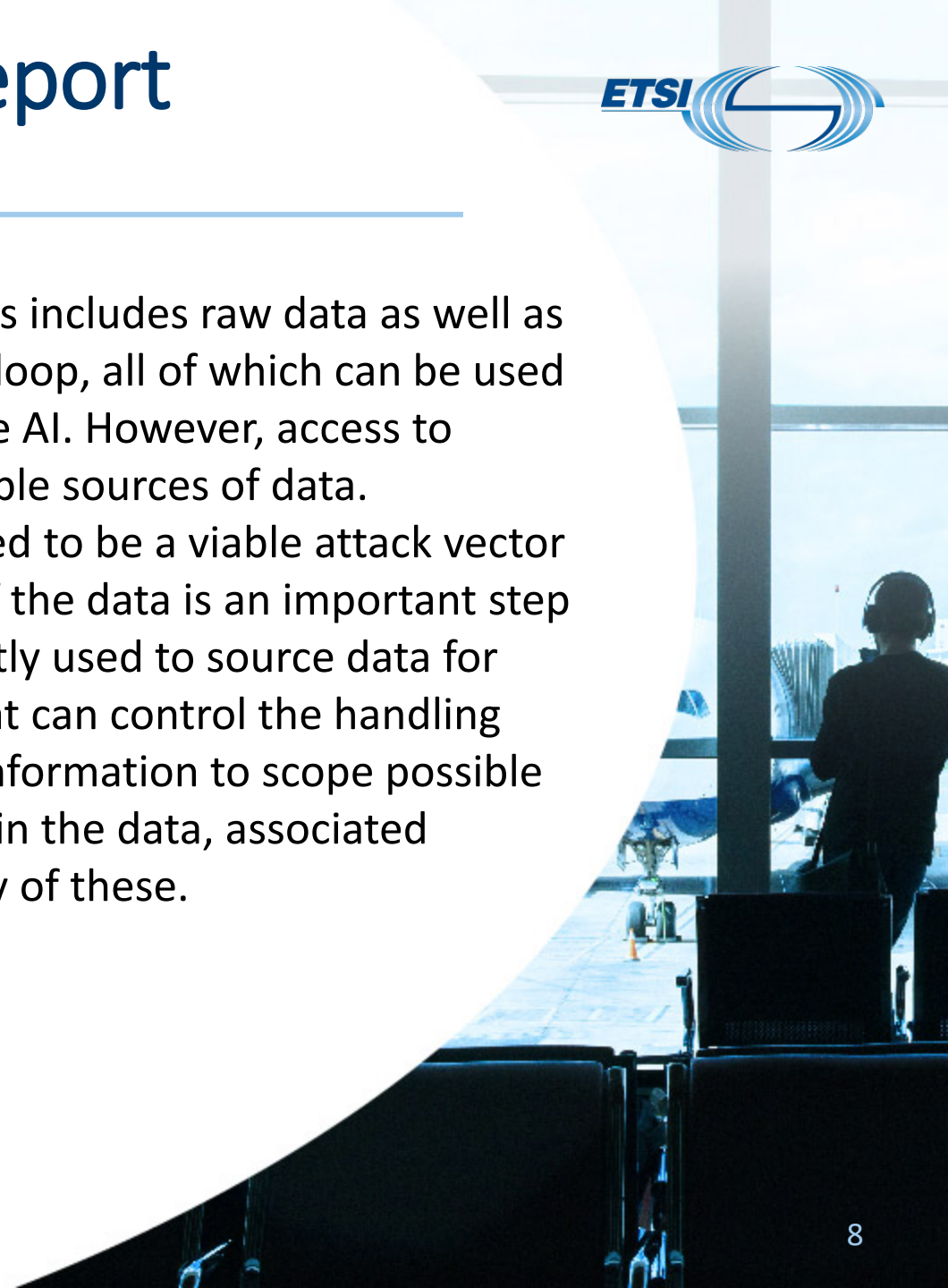
Scope

The purpose of this work item is to define what would be considered an AI threat and how it might differ from threats to traditional systems. The starting point that offers the rationale for this work is that currently, there is no common understanding of what constitutes an attack on AI and how it might be created, hosted and propagated. The AI Threat Ontology deliverable will seek to align terminology across the different stakeholders and multiple industries. This document will define what is meant by these terms in the context of cyber and physical security and with an accompanying narrative that should be readily accessible by both experts and less informed audiences across the multiple industries. Note that this threat ontology will address AI as system, an adversarial attacker, and as a system defender.

GR SAI 002 Data Supply Chain Report (Group Report)

Scope

Data is a critical component in the development of AI systems. This includes raw data as well as information and feedback from other systems and humans in the loop, all of which can be used to change the function of the system by training and retraining the AI. However, access to suitable data is often limited causing a need to resort to less suitable sources of data. Compromising the integrity of training data has been demonstrated to be a viable attack vector against an AI system. This means that securing the supply chain of the data is an important step in securing the AI. This report will summarise the methods currently used to source data for training AI along with the regulations, standards and protocols that can control the handling and sharing of that data. It will then provide gap analysis on this information to scope possible requirements for standards for ensuring traceability and integrity in the data, associated attributes, information and feedback, as well as the confidentiality of these.



GR SAI 005 Mitigation Strategy Report (Group Report)

Scope

This work item aims to summarize and analyze existing and potential mitigation against threats for AI-based systems. The goal is to have guidelines for mitigating against threats introduced by adopting AI into systems. These guidelines will shed light baselines of securing AI-based systems by mitigating against known or potential security threats. They also address security capabilities, challenges, and limitations when adopting mitigation for AI-based systems in certain potential use cases.

Motivation

- Threat Mitigation Report aims to summarize and analyze existing and potential mitigations against threats for AI-based systems.
- It is critical to provide guidelines of threat mitigation against potential /identified threats.
- Threat reports and mitigation reports are complementary to each other
- This work item would summarize known or potential threat mitigations for AI threats and analyze their security capabilities, advantages and suitable scenarios.

GR SAI 006 Role of hardware in SAI (Group Report)

Scope

To prepare a report that identifies the role of hardware, both specialised and general-purpose, in the security of AI. This will address the mitigations available in hardware to prevent attacks (as identified in SAI-005) and address the general requirements on hardware to support SAI (expanding from SAI-004, SAI-002, and SAI-003). In addition this report will address possible strategies to use AI for protection of hardware. The report will also provide a summary of academic and industrial experience in hardware security for AI. In addition, the report will address vulnerabilities or weaknesses introduced by hardware that may amplify attack vectors on AI.

Schedule

- Approved, on publication

Motivation

- A part of this WI involves addressing the concepts of Integrity and Trustworthiness of platforms. With a platform being any computing device (regardless of its architecture or operating system).

- Trustworthiness is required to enable assurance that a system will behave predictably, even under stress. It is difficult to regain trust once lost therefore being able to maintain it it key.

- Aided by platform Integrity assurance which aims to to ensure a verifiable state of the platform, through complete assurance, that under all conditions the systems are correct and reliable. Addressing these are vital for the hardware the AI is running on and making use of them as net contributors to the overall security of the AI system.

Current Work Items

Security Testing of AI (Group Specification)

Scope

The purpose of this work item it to identify objectives, methods and techniques that are appropriate for security testing of AI-based systems. The goal is to have guidelines for testing of AI and AI-based system taking account of the different algorithms. These guidelines will be motivated by the results of the work item "Threat ontology" and quality properties of such systems, new aspect such as testing data for AI in the context of security and addressing challenges when testing AI-based system such as non-determinism and test verdict calculation.

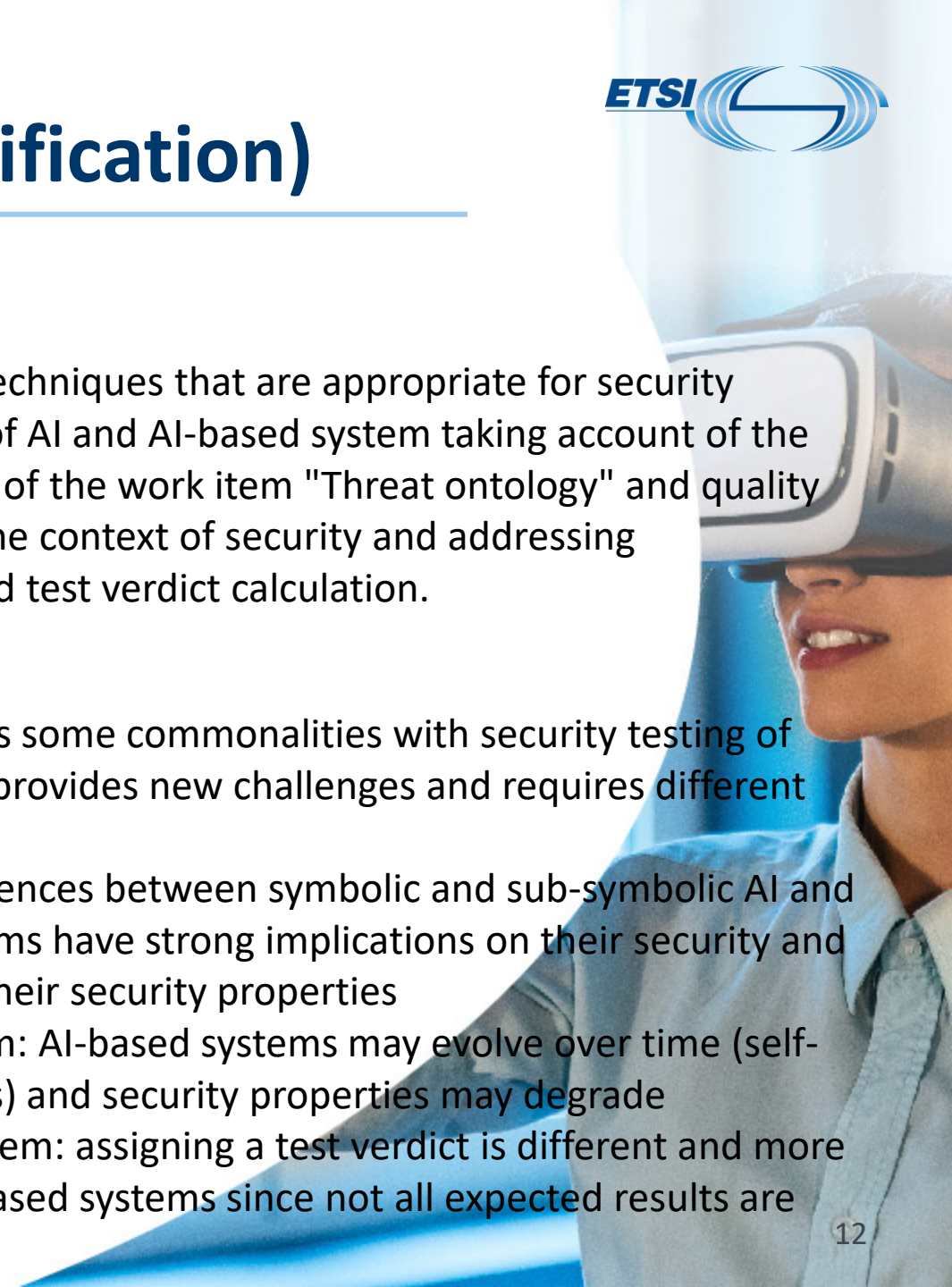
Schedule

TB adoption of WI	**2019/10/23**
Early Draft	2020/01/31
Stable Draft	2021/01/31
TB approval	2022/06/30

Motivation

Security testing of AI has some commonalities with security testing of traditional systems but provides new challenges and requires different approaches, due to:

- significant differences between symbolic and sub-symbolic AI and traditional systems have strong implications on their security and on how to test their security properties
- non-determinism: AI-based systems may evolve over time (self-learning systems) and security properties may degrade
- test oracle problem: assigning a test verdict is different and more difficult for AI-based systems since not all expected results are known a priori



Explicability and transparency of AI processing(Group Report)

Scope

The intent of this work item is to extend from the published work of SAI to address the issues of design of AI platforms (data, algorithms, frameworks) that are able to give assurance of explicability and transparency of decisions. This is intended in part to also consider the impact of issues arising from regulation of AI to address ethics and misuse and to allow independent determination of bias (a light touch). The report will address both intrinsic and post-hoc analysis of AI systems.

Schedule

TB adoption of WI	**2021/06/11**
Early Draft	2021/02/01
Stable Draft	2022/04/30
TB approval	2022/09/30

Privacy aspects of AI/ML systems(Group Report)

Scope

The purpose of this work item is to identify the role of privacy as one of the components of the Security of AI and proceed with the attempt to define Privacy in the context of AI that covers both, safeguarding models and protecting data, as well as the role of privacy-sensitive data in AI solutions. It investigates and addresses the attacks and their associated remediations where applicable, considering the existence of multiple levels of trust affecting the lifecycle of data. Appropriate means to label/protect/anonymize privacy-sensitive data elements during data collection and processing are studied aiming to protect privacy-sensitive data, while limiting AI performance impact. The investigated attack mitigations include Non-AI-Specific (traditional Security/Privacy redresses), AI/ML-specific remedies, pre-emptive remediations (“left of the boom"), and reactive responses to an adversarial activity (“right of the boom”). In addition, the anticipated delivery document will seek to align terminology with existing ETSI SAI ISG documents and studies and will reference previously-studied privacy attacks and remediations (see ETSI GR SAI 004, ETSI GR SAI 002). The anticipated delivery document will also provide a summary of academic and industrial experience in privacy protection for AI.

Schedule

TB adoption of WI	**2021/09/06**
Early Draft	2021/11/08
Stable Draft	2021/12/15
Draft for approval	2022/02/16
Final Draft Published	2022/03/30

Artificial Intelligence Computing Platform Security Framework(Group Report)

Scope

This work item aims to specify a security framework of AI computing platform containing hardware and basic software to protect valuable assets like models and data deployed on AI computing platform when they are used in runtime or stored at rest. The security framework consists of security components in AI computing platform and security mechanisms executed by security components in the platform. By specifying the security framework, AI computing platform can be consolidated against the relevant attack and able to provide security capabilities to facilitate the stakeholders in AI systems to better protect the valuable assets(model/data) on AI computing platform.

Schedule

TB adoption of WI	**2021/10/20**
Early Draft	2022/02/01
Stable Draft	2022/08/31
TB approval	2022/12/31

Traceability of AI Models (Group Report)

Scope

The NWI will study the role of traceability in the challenge of Securing AI and explore issues related to sharing and re-using models across tasks and industries. The scope includes threats, and their associated remediations where applicable, to ownership rights of AI creators as well as to verification of models' origin, integrity or purpose. Mitigations can be non-AI-Specific (Digital Right Management applicable to AI) and AI-specific techniques (e.g. watermarking) from prevention and detection phases. They can be both model-agnostic or model enhancement techniques. Threats and mitigations specific to the collaborative learning setting, implying multiple data and model owners, could be also explored. The NWI will align terminology with existing ETSI ISG SAI documents and studies, and reference/complement previously studied attacks and remediations (ETSI GR SAI 004, ETSI GR SAI 005). It will also gather industrial and academic feedback on traceability and ownership rights protection and model verification (including integrity of model metadata) in the context of AI.

Schedule

TB adoption of WI	**2021/11/16**
Early Draft	2022/01/31
Stable Draft	2022/06/30
TB approval	2022/09/30

Future Work Items

Potential Future Work Items

1. Data Security and Integrity
2. ~~Privacy - enabling or ensuring privacy from the underlying system.~~ -> WI-008 (Privacy aspects of AI/ML systems)
3. Training data: quality, quantity, confidentiality, and labelling
4. ~~Transferability (re-use of models across tasks and industries. Use of watermarking)~~ ->WI-010 (Traceability of AI models)
5. ~~Transparency~~ -> WI-007 (Explicability and transparency of AI processing)
6. ~~Explain–ability (for regulation purposes) (Assurance Issue) (common terminology covering different languages)~~ -> WI-007 (Explicability and transparency of AI processing)
7. Ethics and misuse (While linked to other topics a standalone WI would be inappropriate as is largely non-technical. Though would be considered if it affects other potential topics.
8. Bias and Unintended consequences
9. Data Processing / Machine Learning Life Cycle
10. AI to AI communication (*include transfer learning / knowledge sharing)*
11. AI retraining
12. Prevent manipulation of AI models to give advantage or limit performance depending on Use Case.

Thank you