ETSI ISG Securing Artificial Intelligence

Presented by: **SAI**

For: **ETSI public portal**

04.03.2020

ETSI Securing Artificial Intelligence (SAI) Introduction

•ETSI ISG SAI is the first technology standardization group focusing on securing AI.

•ETSI ISG SAI was officially formed in September 2019; Kickoff meeting on 23rd Oct 2019; Second meeting on 20 Jan 2020; Third meeting will be held on 2-3 April.

•Founding members： NCSC， BT， Huawei UK， Telefonica S.A， C3L

•Current scale： 33 members， 6 Participants, together with European Commission as Counsellor

- The Securing Artificial Intelligence Industry Specification Group (ISG SAI) will 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.

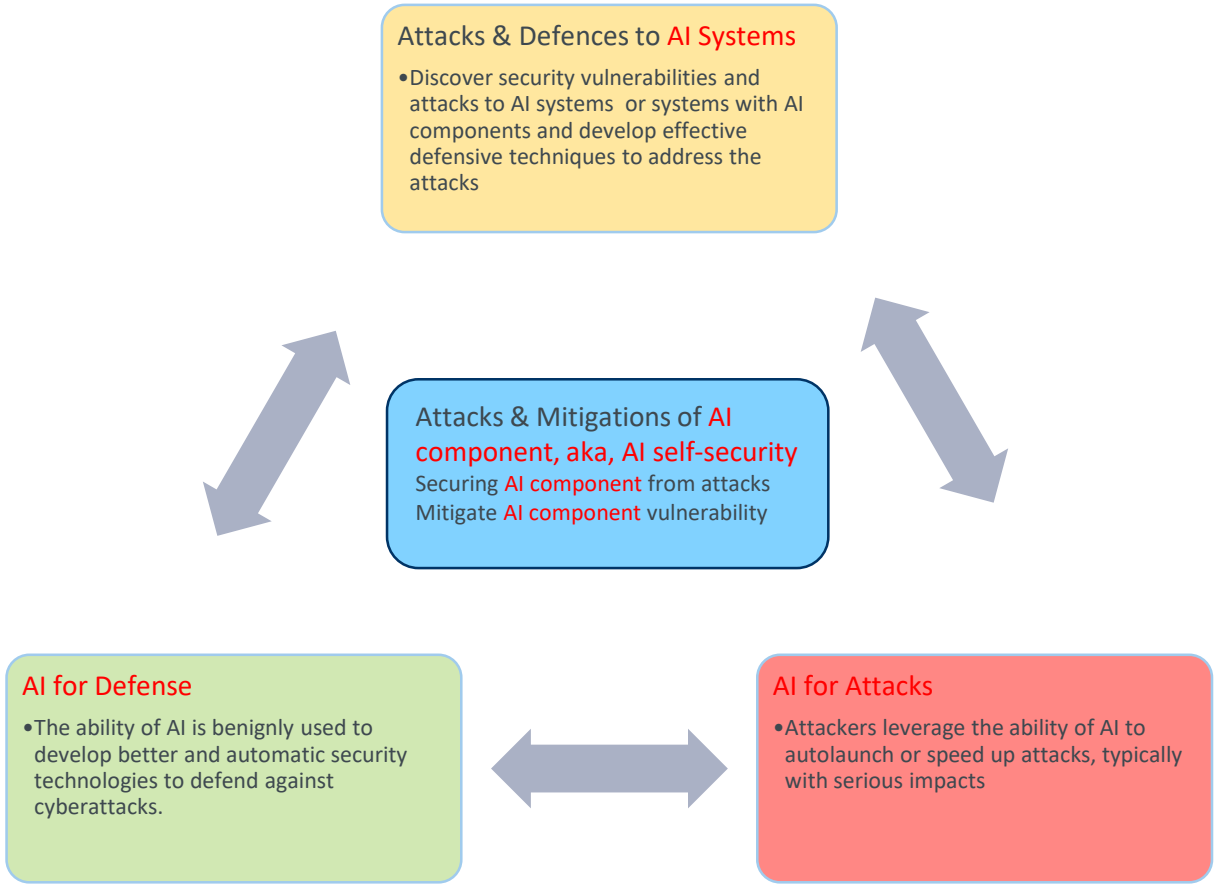
- As a pre-standardisation activity, the ISG SAI is intended to frame the security concerns arising from AI and to build the foundation of a longer-term response to the threats to AI in sponsoring the future developme6t of normative technical specifications.

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

Chairman:	Alex Leadbeater (BT)
Vice-Chairman:	Dr. Kate Reed (NCSC)
Vice-Chairman:	Dr. Tieyan Li (Huawei)
Secretary:	Alexander Cadzow (C3L)
Technical Officer:	Sonia Compans (ETSI)

Five Active Work Items with rapporteurs:

- Securing AI Problem Statement: Philip Mills, Queens University Belfast
- AI Threat Ontology: Scott Cadzow, C3L
- Data Supply Chain: Kate Reed, NCSC
- Mitigation Strategy Report: Hsiao-Ying Lin, Huawei
- Security Testing of AI: Martin Schneider, Fraunhofer FOKUS

Current Work Items

Securing AI Problem Statement (Group Reports)

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.

Schedule

TB adoption of WI	**2020/01/20**
Early Draft	2020/04/20
Stable Draft	2020/07/20
Draft for approval	2020/10/20

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.

AI Threat Ontology (Group Reports)

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.

Schedule

TB adoption of WI	**2019/10/23**
Early Draft	2020/04/03
Stable Draft	2020/07/31
TB approval	2020/09/30

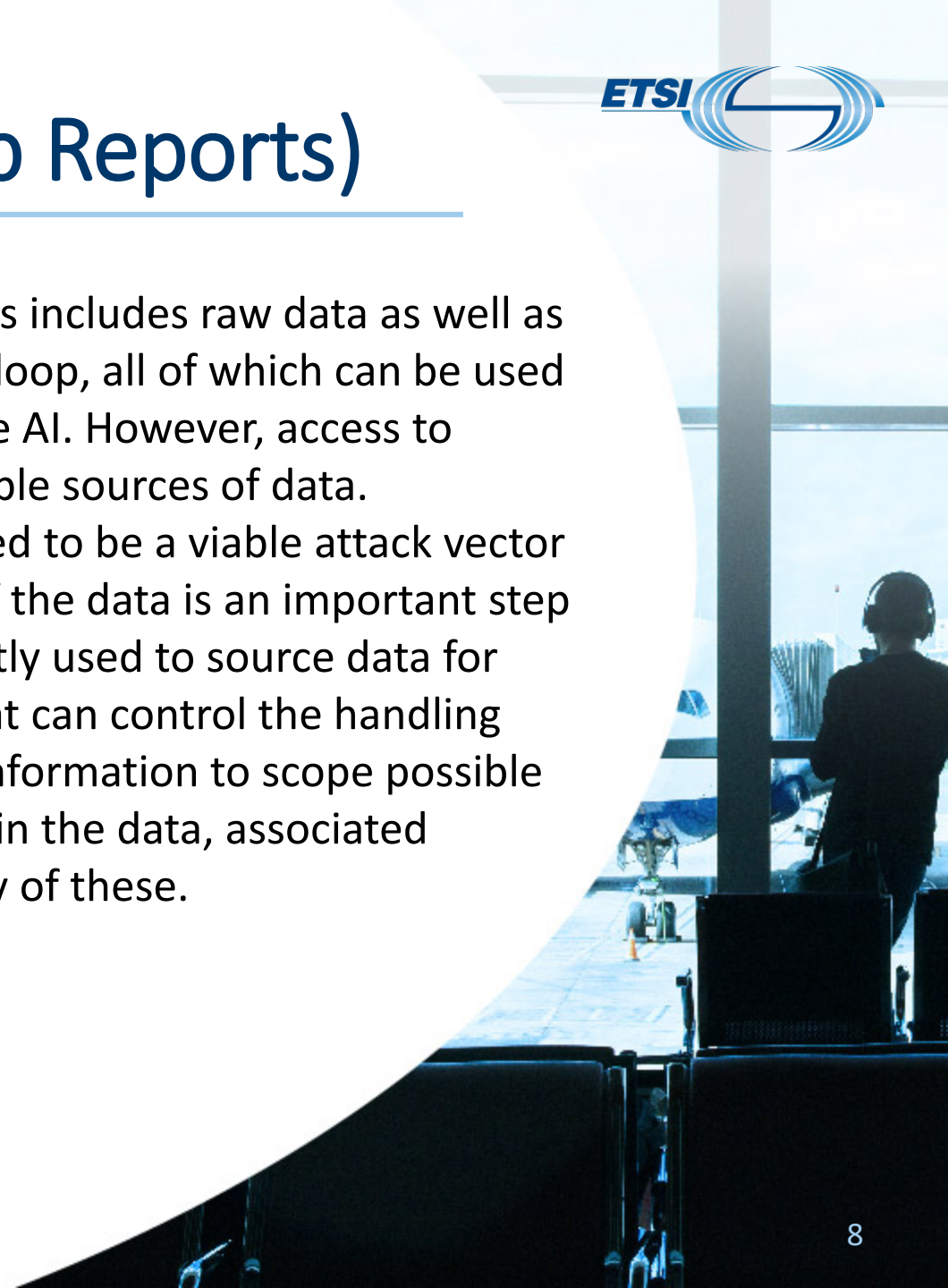
Data Supply Chain Report (Group Reports)

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.

Schedule

TB adoption of WI	**2019/10/23**
Early Draft	2020/04/03
Stable Draft	2020/07/31
TB approval	2020/09/30



Mitigation Strategy Report (Group Reports)

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.

Schedule

TB adoption of WI	**2020/01/20**
Early Draft	2020/06/30
Stable Draft	2020/10/31
Draft for approval	2021/01/31

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.

Security Testing of AI (Group Specifications)

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.

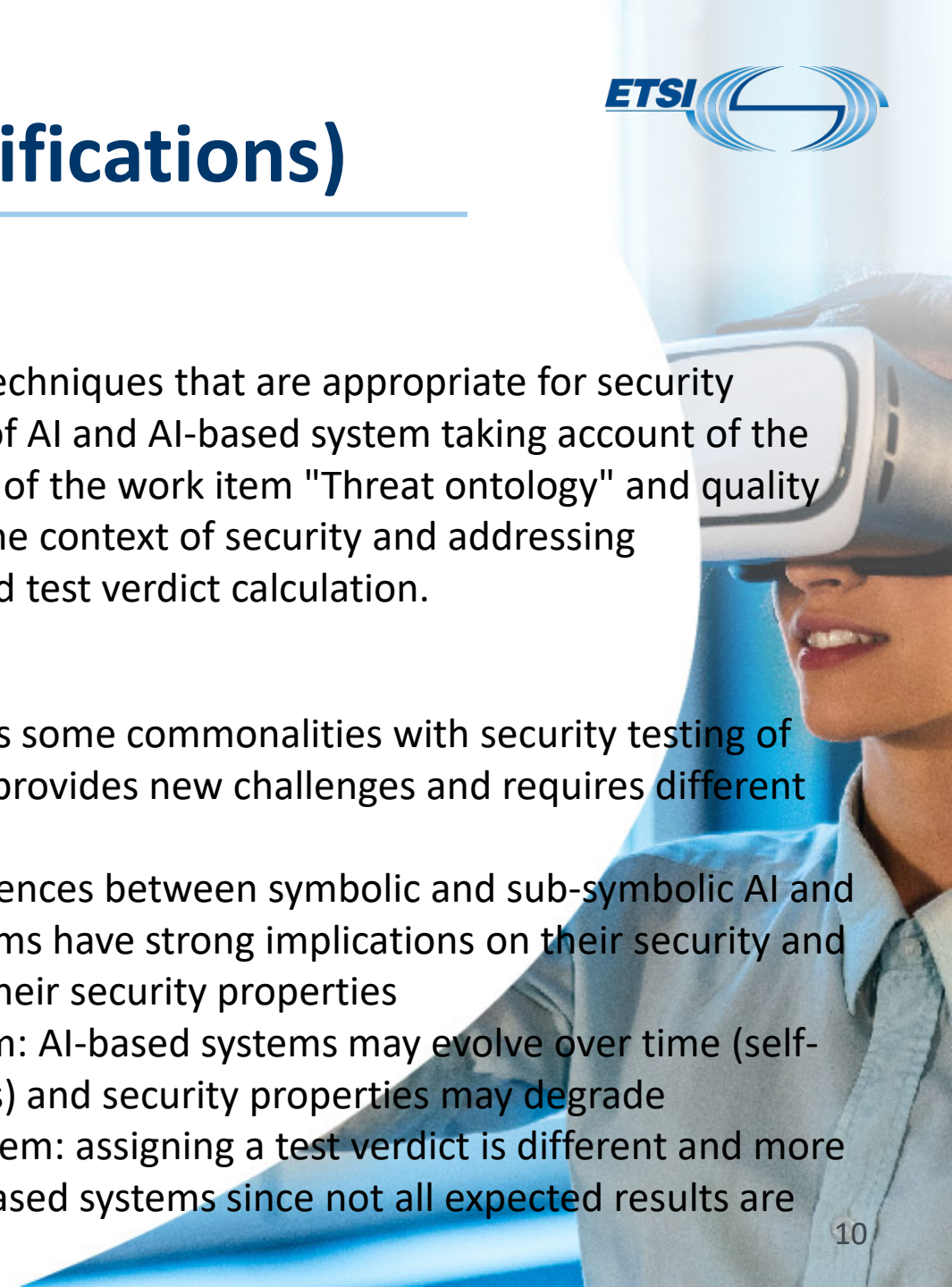
Schedule

TB adoption of WI	**2019/10/23**
Early Draft	2020/01/31
Stable Draft	2020/07/31
TB approval	2020/11/28

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



Q&A

Thank you