TIC COUNCIL

THE INDEPENDENT VOICE OF TRUST

Exploring the Risks and Benefits Presented by Artificial Intelligence applications.

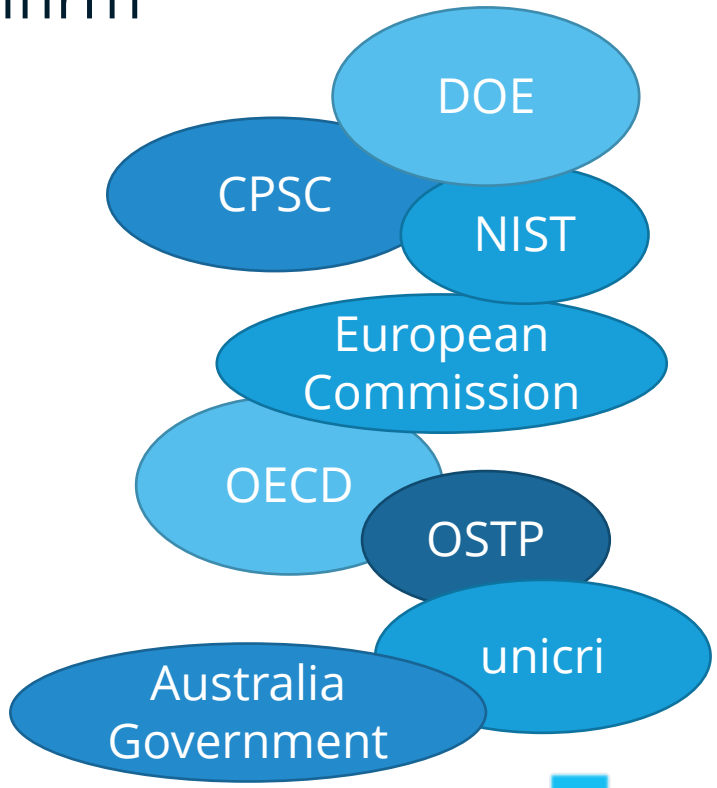
Principles of Artificial Intelligence

"Leaders hoping to shift their posture from hindsight to foresight need to better understand the types of risks they are taking on, their interdependencies, and their underlying causes." – McKinsey Quarterly, Confronting the risks of artificial intelligence, 2019

Principles of Artificial Intelligence

Those elements that are needed to confirm that an AI technology is 'trustworthy –

- Accuracy / Fairness
- Explainability and interpretability
- Privacy
- Reliability / Accountability
- Robustness
- Safety
- Security or resilience to attacks
- Ethics



Accuracy / Fairness

The result produced by the AI technology should be predictable and reliable. (repeatability and reproducibility)

If the data used to train an AI system is biased, the AI will output biased results. (Respect for human rights)

Explainability and interpretability

The ability to explain how the AI technology functions –

- The correlations AI makes between data sets
- Attributes of the data the AI considered
- How the AI reached the conclusion that it did

Privacy

Protecting personally identifiable data

Only using data for the purposes identified

Safety

Having safeguards that protect from potential physical and/or digital harm

Safeguards may include those that prevent

- Disabling of the safety device
- The introduction of a hazard

Safeguards may also include alerts or notices to the customer/user

Security or Resilience to Attacks

Having safeguards in place to protect from cybersecurity risks

Having safeguards to prevent malicious data from being introduced to the AI

Examples:

- Application security
- Cloud security
- IoT security
- Vulnerability assessments

Reliability / Accountability

Auditable AI

The ability to confirm through other means that the data evaluated would lead to the outcome produced by the AI

Robustness

Ability of the AI to perform as expected, even while undergoing rigorous testing in changing conditions meant to challenge the AI or cause it to perform not as intended.

Ethics

The AI should not be used to cause foreseeable or unintentional harm – a loss of trust and negative impacts on social well being.

Examples

- Used to discriminate against protected grouped
- Used to mislead the public (e.g., AI generated images)
- Used to manipulate events, views, or communications

Jacques Kruse Brandao
Global Head of Advocacy
SGS

Risk Key Areas

- **Human agency and oversight**
  Including fundamental rights, human agency and human oversight
- **Technical robustness and safety**
  Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility
- **Privacy and data governance**
  Including respect for privacy, quality and integrity of data, and access to data
- **Transparency**
  Including traceability, explainability and communication
- **Diversity, non-discrimination and fairness**
  Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation
- **Societal and environmental wellbeing**
  Including sustainability and environmental friendliness, social impact, society and democracy
- **Accountability**
  Including auditability, minimization and reporting of negative impact, trade-offs and redress

https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html

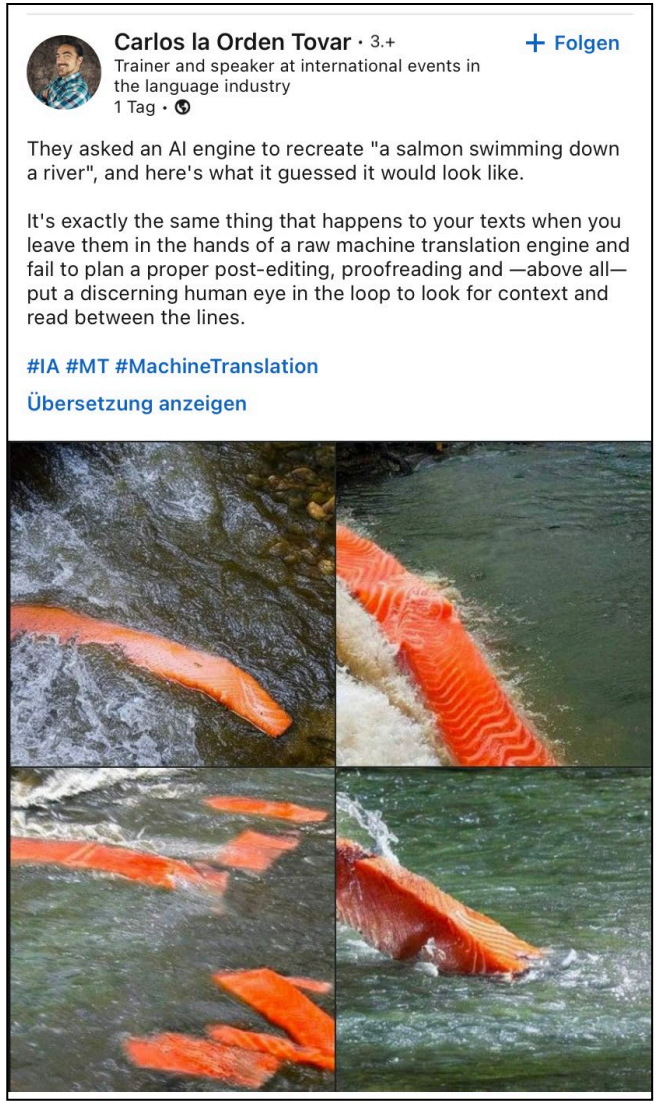
Challenges in AI

- Standards on quality of data: „Salmon swimming in a river“
- Deep Fakes like audio & voice imitating CEOs asking for money transfer
- Safety in automotive driving
- Wrong decision in medical diagnostic
- …

How to enable AI Applications while we are facing those issues?

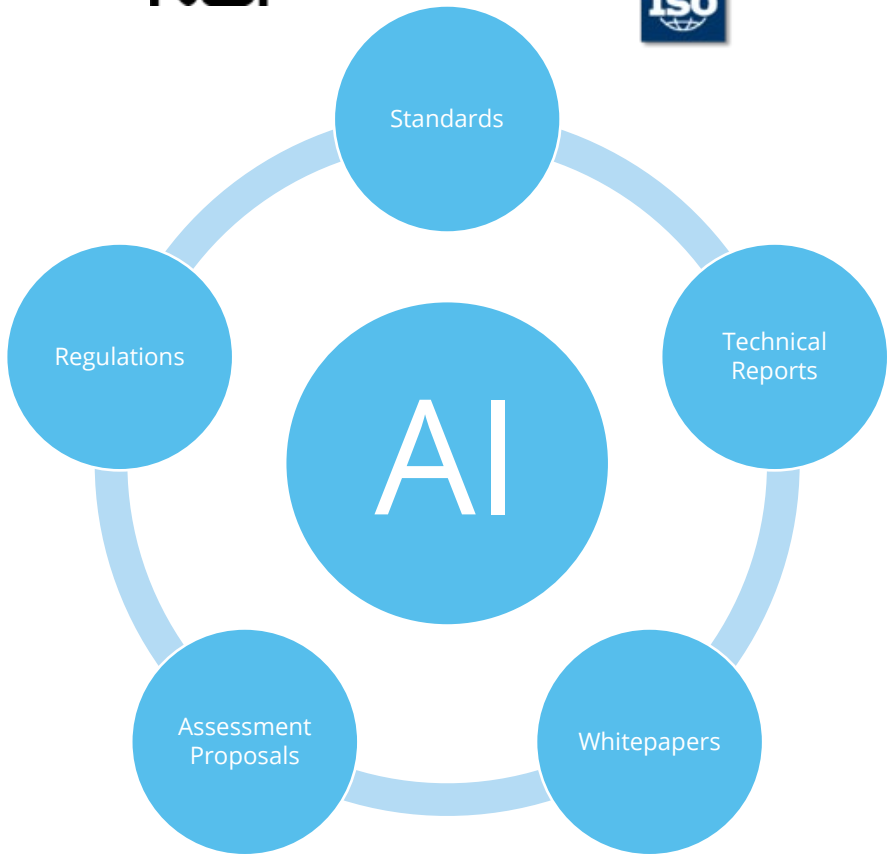
⇒ Need for assessment from various angles based on
  - **Framework that everybody understands**
  - **Supporting legislation**
  - **Standards**
  - **Test Methodologies**

to generate **trust** to the user.



AI Publication Landscape

OECD maintains a live repository of over **700 AI policy initiatives** from 60 countries, territories and the EU



AI Regulation or regulation-like activities

- **EU**:
  - AI Act

- **US**:
  - FDA: Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning - Based Software as a Medical Device
  - FTC: Aiming for truth, fairness, and equity in your company's use of AI
  - NIST AI Rim Framework

- **China**:
  - Internet Information Service Algorithmic Recommendation Management Provisions

EU in the lead with the AI Act

Scope: Trust in AI

Proposal for Regulation laying down harmonized rules on AI

- **Minimum requirements** necessary to address risks and problems linked to AI.
  It requires conformity assessments for high-risk applications
- **Limited number of AI applications creating unacceptable risk**, e.g., social scoring or remote biometric identification
- Number of **high-risk applications** that, while not prohibited, should be subject to regulatory requirements:
  - use of **facial recognition technology** in the area of law enforcement (guideline in work)
  - discussion is about **substantial broadening** of the list of high-risk AI applications
- Excluded from the AI Act: **medical devices, civil aviation, vehicles, maritime** and **rail sector**
  -> justification: current sectoral rules are already stricter
- AI Act has **references** to the **CSA, Machinery Directive, GDPR** or the protection of **fundamental rights** and **safety**

Depart from the AI Act the EC also drafted the **AI Liability Directive** as well as the new **Product Liability Directive.**

Standards

Standards and associated **test methodologies** need jointly to be defined related to

Risk Assessment & Risk Classification	-> risk levels, assurance levels
Fairness / Bias	-> *acceptable level of bias?*
Autonomy / Control	-> *Human in the loop*
Transparency	-> data, system, business model
Explainability	-> *description of development steps*
Performance / Reliability	-> *Monitoring and corrective measures to the output*
Functional Safety	-> *coherence with Machinery Directive?*
Robustness / Cybersecurity	-> *Cyber Resilience Act defines essential requirements*
Privacy	-> *First GDPR Certification scheme approved by the EDPB!*
Data Quality / Traceability	-> *documentation incl. attributes of the data sets?*

First proposals from different initiatives are good input already:

Standards & Certification related to AI

Over 20 new standards are already in development at various standardization organizations.
Several initiatives and publications across the globe aiming for establishing methodologies to build trust in AI systems.

Standards in development:

- **ETSI** GR SAI 005 : “Securing Artificial Intelligence”
- **CEN/CENELEC**
  - Standards dedicated to AI (ISO/IEC JTC1/SC42)
  - Standards dedicated to Cybersecurity (ISO SC27)
  - Standards dedicated to Biometry (ISO SC37)
- **ISO/IEC JTC 1 / SC42**
  - ISO/IEC TR 24029-1 “ Assessment of the robustness of neuronal networks”
  - ISO/IEC 24028 : “Overview of trustworthiness in Intelligence artificial”
  - ISO 38057
- **IEEE** 700x-Series
- **DIN Standardization Roadmap AI**
  - AG 2 – Testing and Certification
  - -> proposal planned to be published in January 2023

European Commission: Standardization Request on AI:
Deadline for the adoption by the ESOs: 31/10/2024

Expected Timeline:

2022: Development of Criteria, Test Methodologies and Certification schemes for high-risk applications
2023: Validation of Criteria against Use Cases / Development of missing standards
2024: Harmonized Standards

Audability – what can be assessed/ certified?

Process Certification

Design process
Development processes
Data evaluation processes

Product Certification

Assessment of AI functionality
Conformity assessment against defined specifications
Maintenance process: Re-Assessment after changes of underlying data or algorithms

Certification of People/ Organizations

People involved development
People involved in using AI
Organization Maturity Level

Standardization & verification of

data quality assurance,
supply chain
training process (certifiable training)
evaluation process
final decision logic (neural networks)

Verification of the functional safety

Evaluation/selection of AI (sub-)systems
Risk Analysis of each AI (sub-)systems
Verification and validation of the implementation of each AI (sub-)System
Evaluation, verification and validation of each modulation of each AI (sub-)System
Iterative process until decommissioning

Verification and validation of the Adversarial training (certifiable defenses to improve the robustness

Failure or malfunction

Proper operation of an AI system

has to be guaranteed under realistically defined boundary conditions

Quantifying the risk for system failure,

i.e. the cost of failure multiplied with the probability of failure

Whole life cycle of AI systems need to be considered

*-> To raise the **Maturity Level of supplier of AI systems** we need **Consulting**, **Training** and **Assessment** before any certification will start.*

TIC COUNCIL

Follow us online

@TICCouncil

TIC Council

Wikipedia page:
Testing, inspection and certification

TIC-Council.org