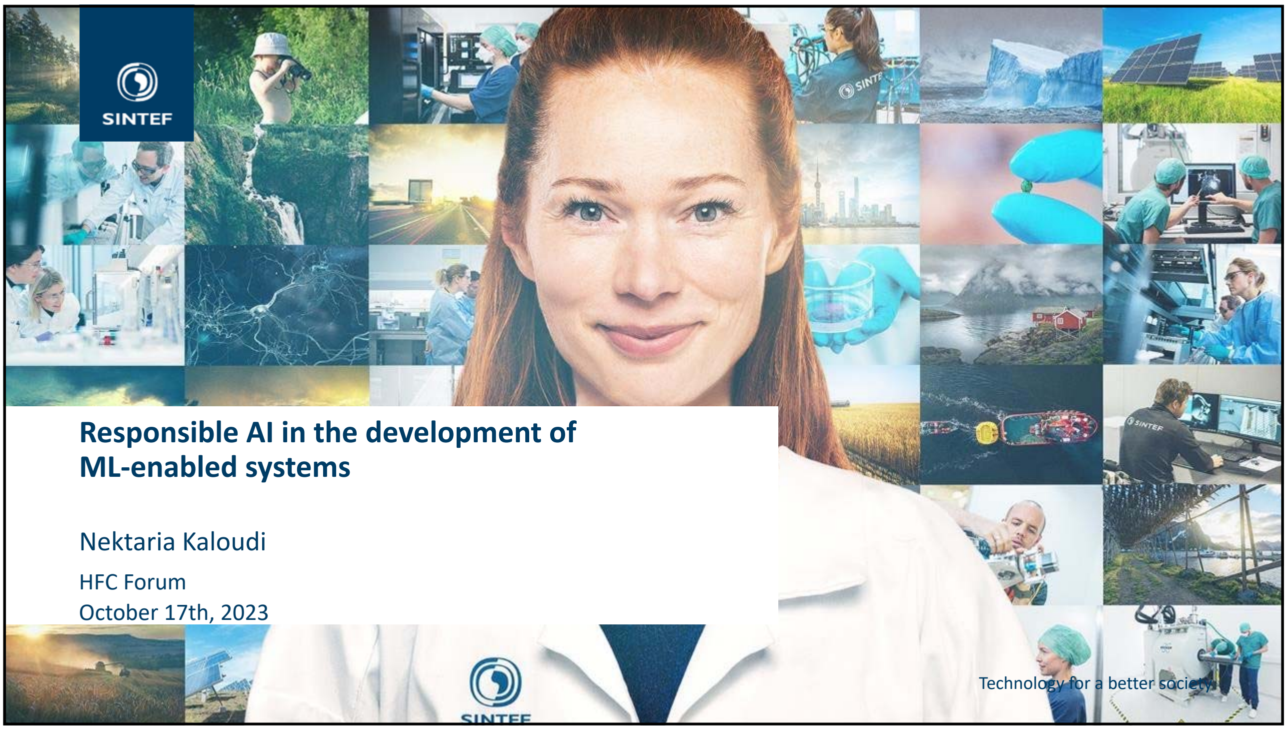


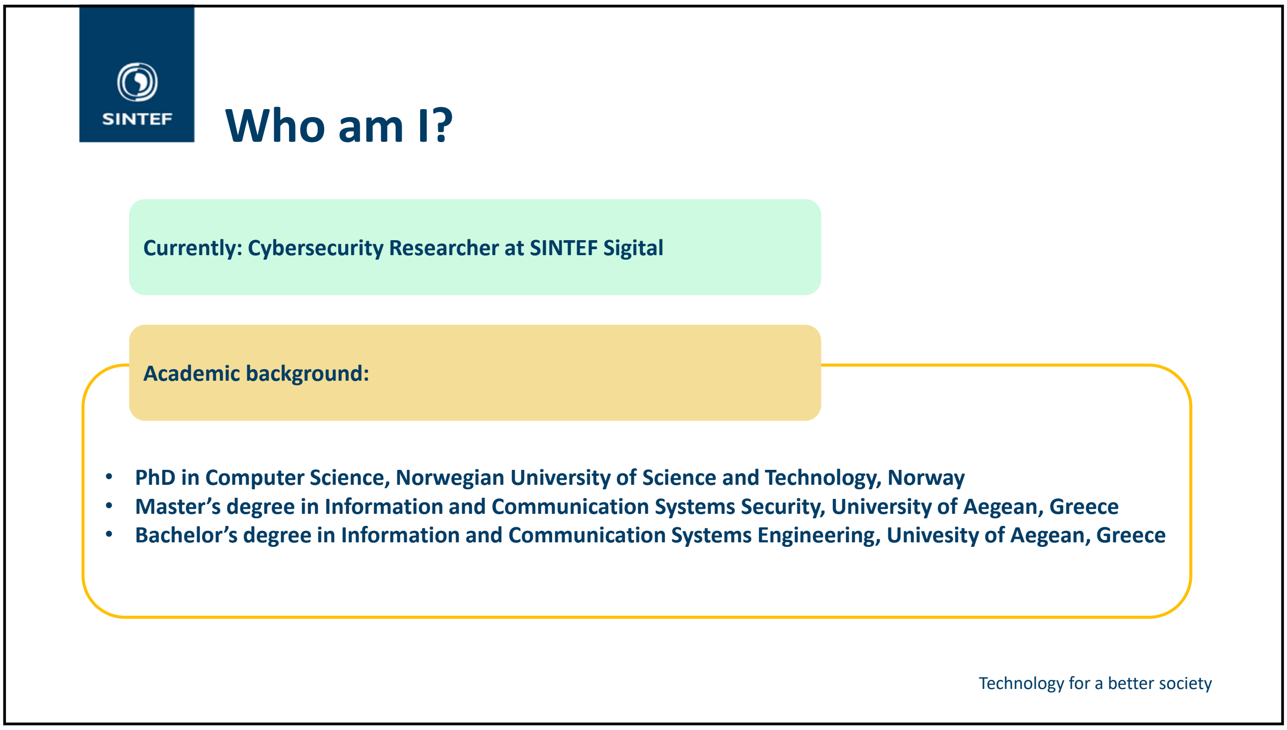
Responsible AI in the development of ML-enabled systems

Nektaria Kaloudi

HFC Forum
October 17th, 2023

Technology for a better society

1



Who am I?

Currently: Cybersecurity Researcher at SINTEF Sigital

Academic background:

- **PhD in Computer Science, Norwegian University of Science and Technology, Norway**
- **Master's degree in Information and Communication Systems Security, University of Aegean, Greece**
- **Bachelor's degree in Information and Communication Systems Engineering, Univesity of Aegean, Greece**

Technology for a better society

2

SINTEF

Today’s talk

- AI and ML
- Examples of real-world incidents
- Responsible AI
- Principles
- The case of cybersecurity
- Ways of moving forward

Technology for a better society

3

SINTEF

AI and ML

- Artificial intelligence (AI) has been first defined by John McCarthy in 1955 as “*the science and engineering of making intelligent machines*”
- Machine learning (ML) is the type of AI that learns from data and turns into predictions or decisions

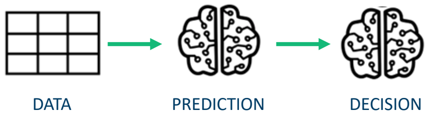
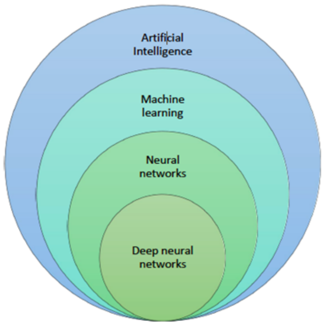


Figure 1. Relationship between AI and ML

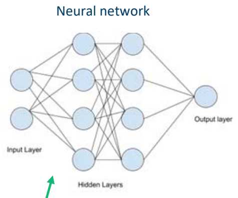


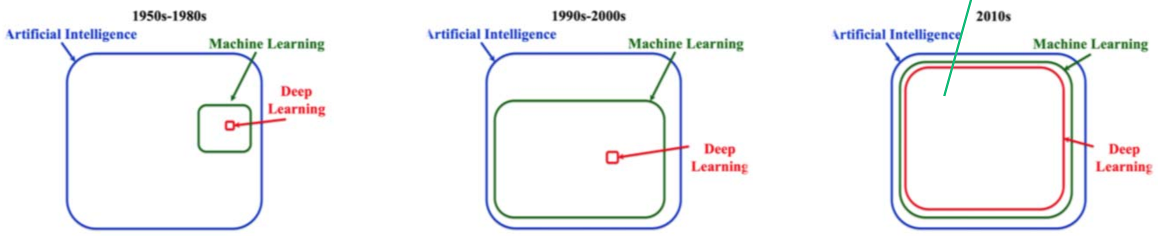
CEPS Task Force Report “*Artificial Intelligence and Cybersecurity*”, 2021

Technology for a better society

4

AI and its terminology





Algorithm + Data + Computation

Melanie Mitchell “*Artificial Intelligence: A Guide for Thinking Humans*”, 2019
Gary McGraw “*Security Engineering for Machine Learning*”, 2022

Technology for a better society

5

Real-world incidents

6

Self-driving cars

Sources: CNN "Another Tesla reportedly using Autopilot hits a parked police car", 2021
The Guardian "Tesla that crashed into police car was in autopilot mode", 2018

Technology for a better society

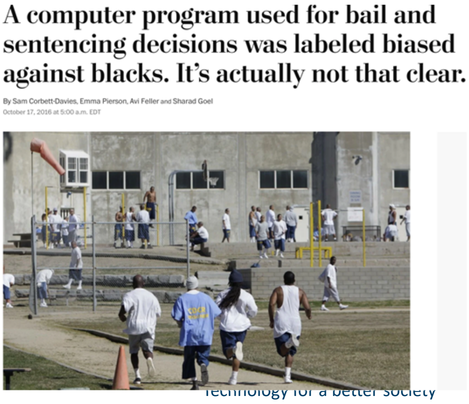
7

Fairness

SCIENCE & TECH | SOCIAL POLICY
The future of workers' rights in the AI age

Houston teachers fired by automated system

Amazon's face recognition falsely matched 28 members of Congress with mugshots



A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

By Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel
October 17, 2016 at 5:00 a.m. EDT

Technology for a better society

8

Responsible AI

9

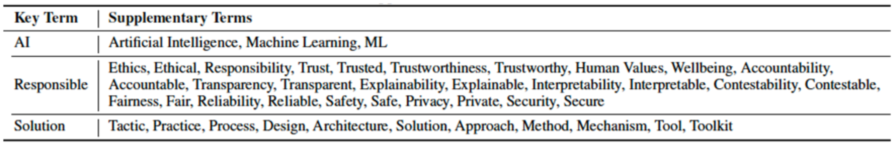
SINTEF

Terminology

How can we ensure that AI systems are developed and adopted in a responsible way?

*According to Dignum[1], **Responsible AI** is about being responsible for the power that AI brings.*

Also commonly referred to as "*trustworthy AI*" or "*ethical AI*", with a common goal to promote development, deployment, and use of AI systems that have a positive impact on individuals, society, and environment while minimizing associated risks.



Key Term	Supplementary Terms
AI	Artificial Intelligence, Machine Learning, ML
Responsible	Ethics, Ethical, Responsibility, Trust, Trusted, Trustworthiness, Trustworthy, Human Values, Wellbeing, Accountability, Accountable, Transparency, Transparent, Explainability, Explainable, Interpretability, Interpretable, Contestability, Contestable, Fairness, Fair, Reliability, Reliable, Safety, Safe, Privacy, Private, Security, Secure
Solution	Tactic, Practice, Process, Design, Architecture, Solution, Approach, Method, Mechanism, Tool, Toolkit

Technology for a better society

[1] *Dignum, Virginia. Responsible artificial intelligence: how to develop and use AI in a responsible way. Springer, 2019.*

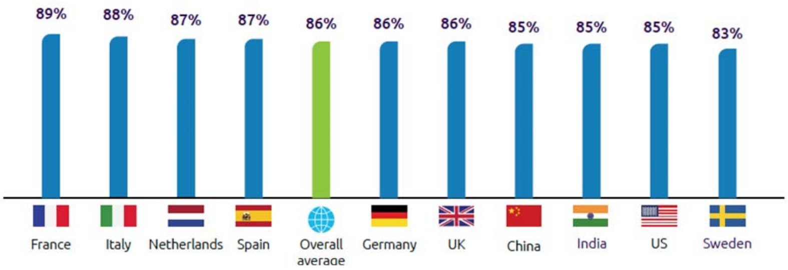
10

SINTEF

Why Responsible AI?

- Advancements in AI differ from those in other technologies due to the **rapid pace of innovation** and their **proximity to human intelligence**, which impacts us on both personal and societal levels.
- Nine out of ten organizations across countries have encountered ethical issues resulting from the use of AI

In the last 2-3 years, have the below issues resulting from the use and implementation of AI systems, been brought to your attention? (percentage of executives, by country)

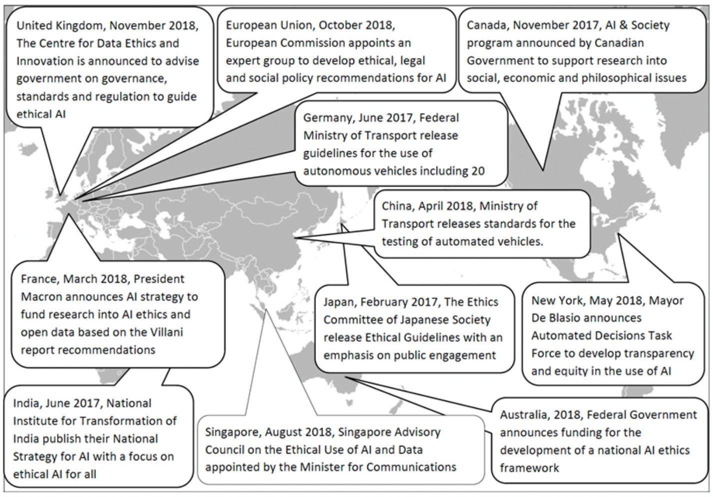


Source: Capgemini Research Institute "Why addressing ethical questions in AI will benefit organizations", 2019

11

SINTEF

Worldwide efforts



United Kingdom, November 2018, The Centre for Data Ethics and Innovation is announced to advise government on governance, standards and regulation to guide ethical AI

European Union, October 2018, European Commission appoints an expert group to develop ethical, legal and social policy recommendations for AI

Canada, November 2017, AI & Society program announced by Canadian Government to support research into social, economic and philosophical issues

Germany, June 2017, Federal Ministry of Transport release guidelines for the use of autonomous vehicles including 20

China, April 2018, Ministry of Transport releases standards for the testing of automated vehicles.

France, March 2018, President Macron announces AI strategy to fund research into AI ethics and open data based on the Villani report recommendations

Japan, February 2017, The Ethics Committee of Japanese Society release Ethical Guidelines with an emphasis on public engagement

New York, May 2018, Mayor De Blasio announces Automated Decisions Task Force to develop transparency and equity in the use of AI

India, June 2017, National Institute for Transformation of India publish their National Strategy for AI with a focus on ethical AI for all

Singapore, August 2018, Singapore Advisory Council on the Ethical Use of AI and Data appointed by the Minister for Communications

Australia, 2018, Federal Government announces funding for the development of a national AI ethics framework

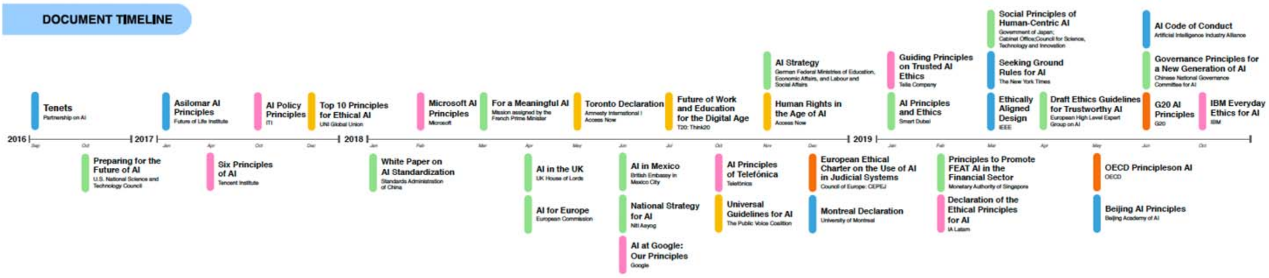
Source: Dawson, D., Schleiger, E., Horton, J., McLaughlin, J., Robinson, C., Quezada, G., ... & Hajkowicz, S. (2019). Artificial intelligence: Australia's ethics framework-a discussion paper.

Technology for a better society

12

SINTEF

Principled Artificial Intelligence



A map of ethical and rights-based approaches to principles for AI

Nature of Actors

- 🟨 Civil Society
- 🟩 Government
- 🟧 Inter-governmental Organization
- 🟦 Multistakeholder
- 🟪 Private Sector

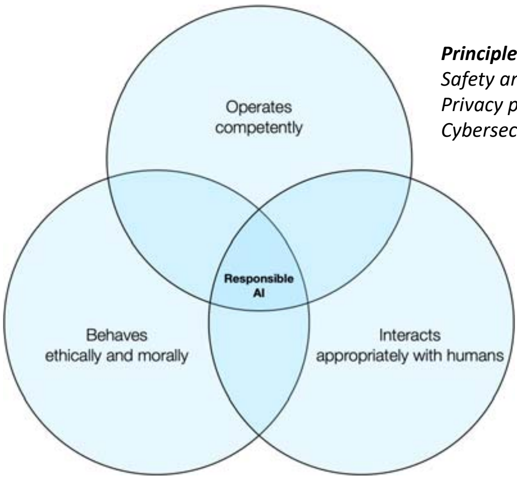
Source: Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication, (2020-1).

SINTEF

Main pillars of responsible AI

What would constitute a responsible autonomous agent?

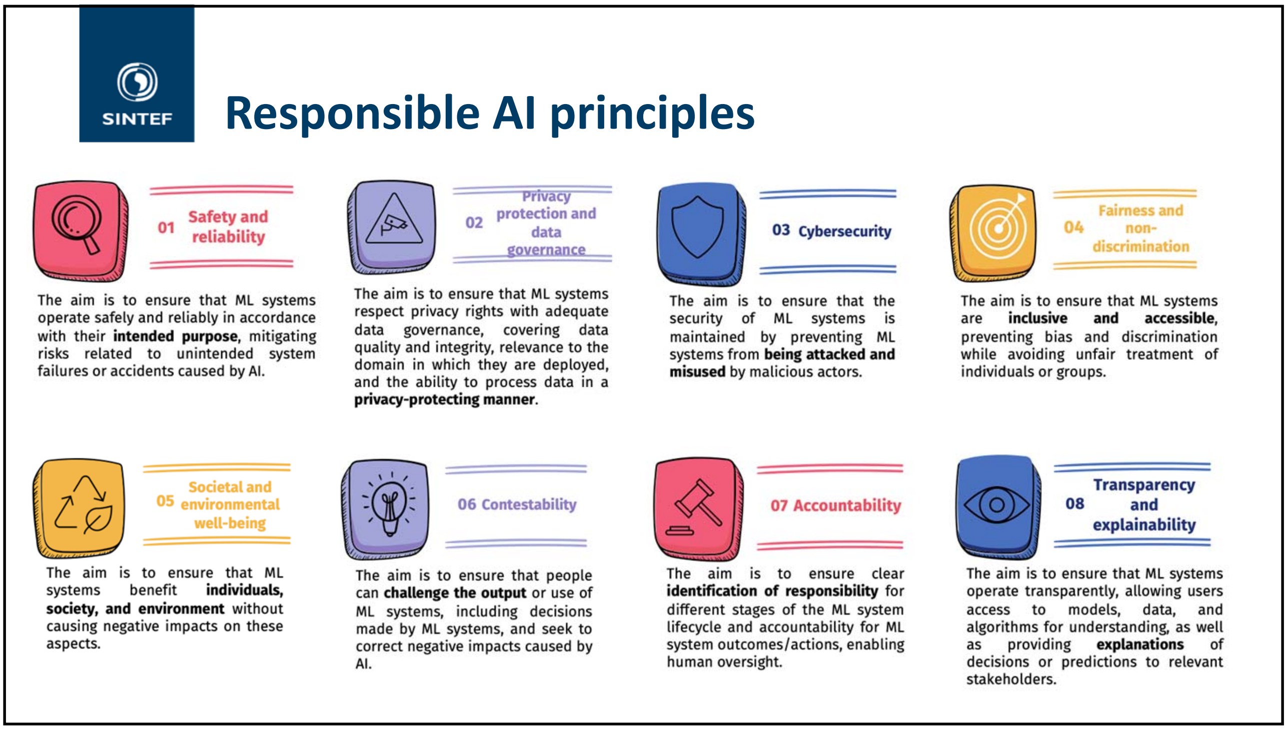
What we need to care about in order to have Responsible AI


Operates competently
Responsible AI
Behaves ethically and morally
Interacts appropriately with humans

***Principles**:*
Safety and reliability
Privacy protection and data governance
Cybersecurity

***Principles**:*
Fairness and non-discrimination
Societal and environmental well-being

***Principles**:*
Accountability
Transparency and explainability
Contestability



SINTEF

Responsible AI principles

01 Safety and reliability

The aim is to ensure that ML systems operate safely and reliably in accordance with their **intended purpose**, mitigating risks related to unintended system failures or accidents caused by AI.

02 Privacy protection and data governance

The aim is to ensure that ML systems respect privacy rights with adequate data governance, covering data quality and integrity, relevance to the domain in which they are deployed, and the ability to process data in a **privacy-protecting manner**.

03 Cybersecurity

The aim is to ensure that the security of ML systems is maintained by preventing ML systems from **being attacked and misused** by malicious actors.

04 Fairness and non-discrimination

The aim is to ensure that ML systems are **inclusive and accessible**, preventing bias and discrimination while avoiding unfair treatment of individuals or groups.

05 Societal and environmental well-being

The aim is to ensure that ML systems benefit **individuals, society, and environment** without causing negative impacts on these aspects.

06 Contestability

The aim is to ensure that people can **challenge the output** or use of ML systems, including decisions made by ML systems, and seek to correct negative impacts caused by AI.

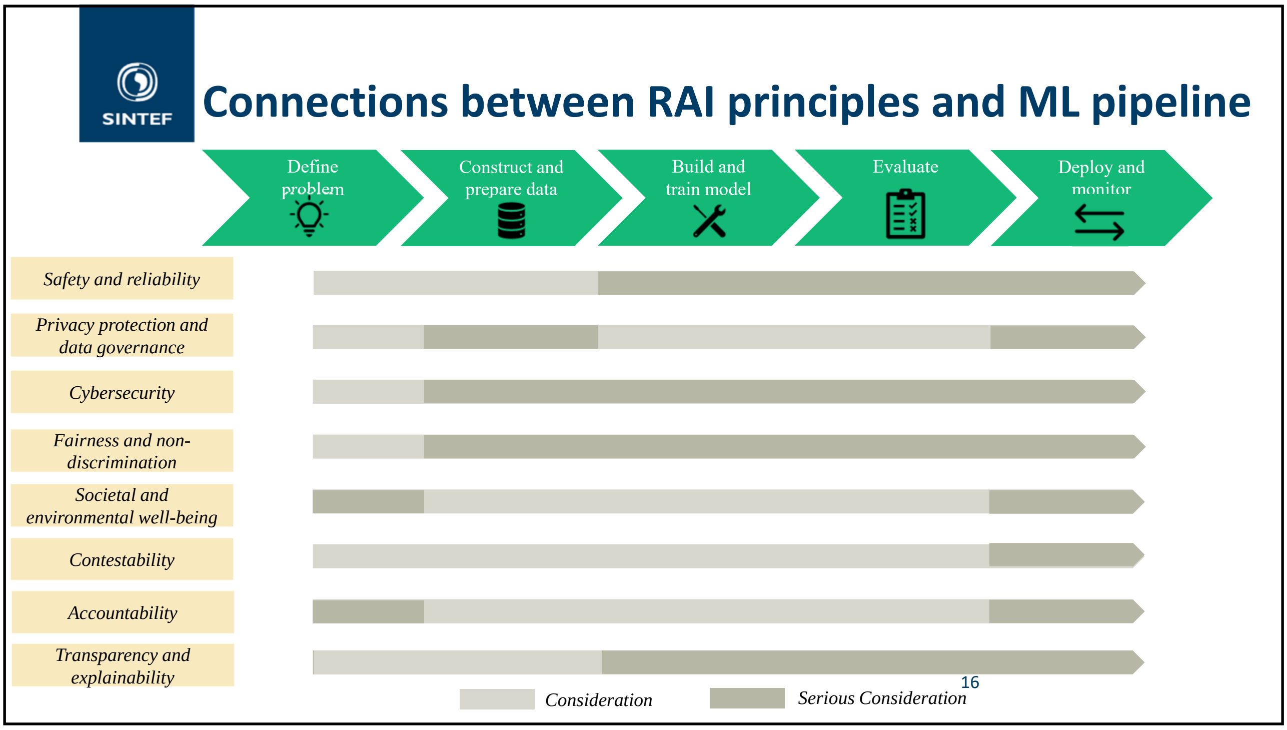
07 Accountability

The aim is to ensure clear **identification of responsibility** for different stages of the ML system lifecycle and accountability for ML system outcomes/actions, enabling human oversight.

08 Transparency and explainability

The aim is to ensure that ML systems operate transparently, allowing users access to models, data, and algorithms for understanding, as well as providing **explanations** of decisions or predictions to relevant stakeholders.

15



SINTEF

Connections between RAI principles and ML pipeline

16

Implementing RAI principles

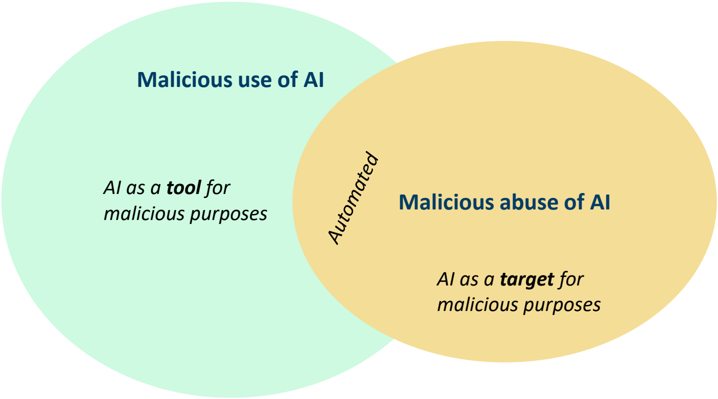
the case of cybersecurity

17

SINTEF

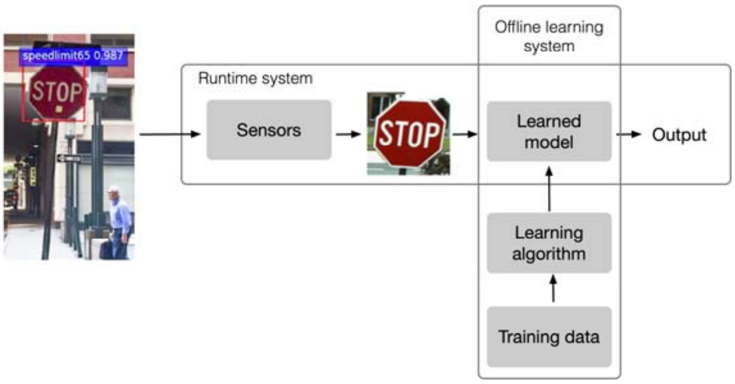
The cybersecurity principle

Malicious AI can be seen through the lens of **malicious use of AI** and **malicious abuse of AI**


Malicious use of AI
*AI as a **tool** for malicious purposes*
Automated
Malicious abuse of AI
*AI as a **target** for malicious purposes*

18

Malicious abuse of AI



Offline learning system

Runtime system

Sensors

Learned model

Output

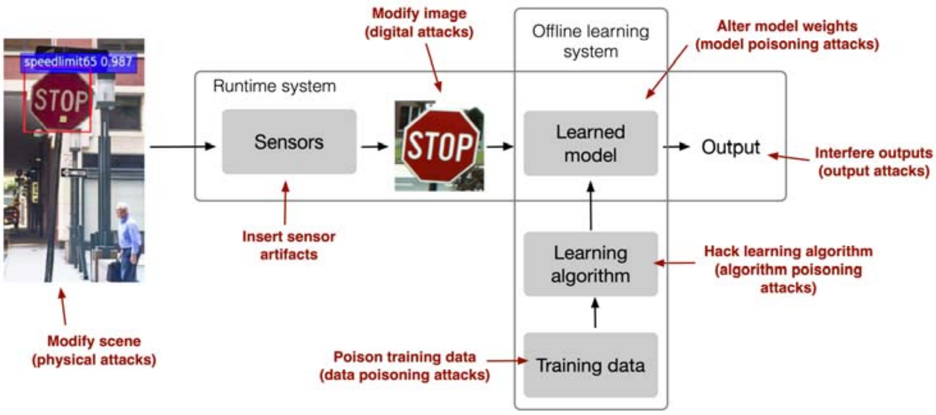
Learning algorithm

Training data

Technology for a better society

19

Threat model: what can the attacker do?



Modify image (digital attacks)

Offline learning system

Alter model weights (model poisoning attacks)

Runtime system

Sensors

Learned model

Output

Interfere outputs (output attacks)

Insert sensor artifacts

Learning algorithm

Hack learning algorithm (algorithm poisoning attacks)

Modify scene (physical attacks)

Poison training data (data poisoning attacks)

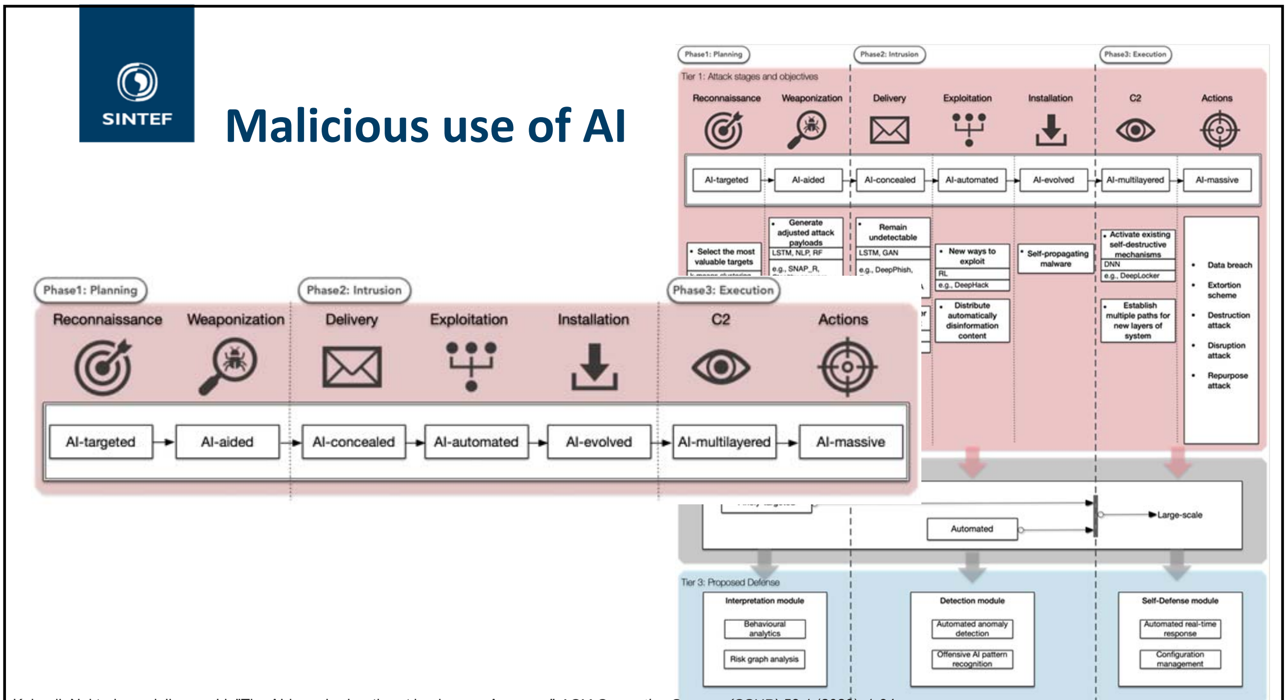
Training data

Technology for a better society

20

SINTEF

Malicious use of AI



Phase1: Planning

Tier 1: Attack stages and objectives

Reconnaissance

Weaponization

Phase2: Intrusion

Delivery

Exploitation

Installation

Phase3: Execution

C2

Actions

AI-targeted → AI-aided → AI-concealed → AI-automated → AI-evolved → AI-multilayered → AI-massive

- Select the most valuable targets

- Generate adjusted attack payloads

LSTM, NLP, RF

e.g., SNAP_R,

- Remain undetectable

LSTM, GAN

e.g., DeepPhish,

- New ways to exploit

RL

e.g., DeepHack

- Distribute automatically disinformation content

- Self-propagating malware

- Activate existing self-destructive mechanisms

DNN

e.g., DeepLocker

- Establish multiple paths for new layers of system

- Data breach
- Extortion scheme
- Destruction attack
- Disruption attack
- Repurpose attack

Automated

Large-scale

Tier 3: Proposed Defense

Interpretation module

Behavioural analytics

Risk graph analysis

Detection module

Automated anomaly detection

Offensive AI pattern recognition

Self-Defense module

Automated real-time response

Configuration management

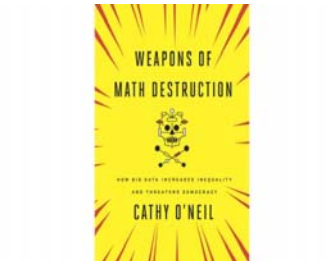
Kaloudi, Nektaria, and Jingyue Li. "The AI-based cyber threat landscape: A survey." *ACM Computing Surveys (CSUR)* 53.1 (2020): 1-34.

SINTEF

Cybersecurity for AI

A narrow and traditional scope: the protection against attacks across the lifecycle of an AI system

- Improving the lack of robustness and vulnerabilities of AI models
- Defending AI systems from attacks (e.g., manipulation of data used in AI systems, attacks against AI-powered CPS, data poisoning, environment variations can be caused on the data)



Algorithms Are Opinions Embedded in Code

Technology for a better society

SINTEF

Cybersecurity for AI

A broad and extended scope: supporting with trustworthiness features

- Credible and reliable training datasets, dealing with bias – *fairness principle*
- Algorithmic validation and verification – *safety and reliability principle*
- Data protection and privacy in the context of AI systems – *privacy and data governance principle*
- Clear identification of responsibility for different stages of ML system lifecycle – *accountability principle*
- Explanations of decisions made by AI systems – *transparency and explainability principle*
- Ensure that people can challenge the output of ML systems – *the contestability principle*

Technology for a better society

23

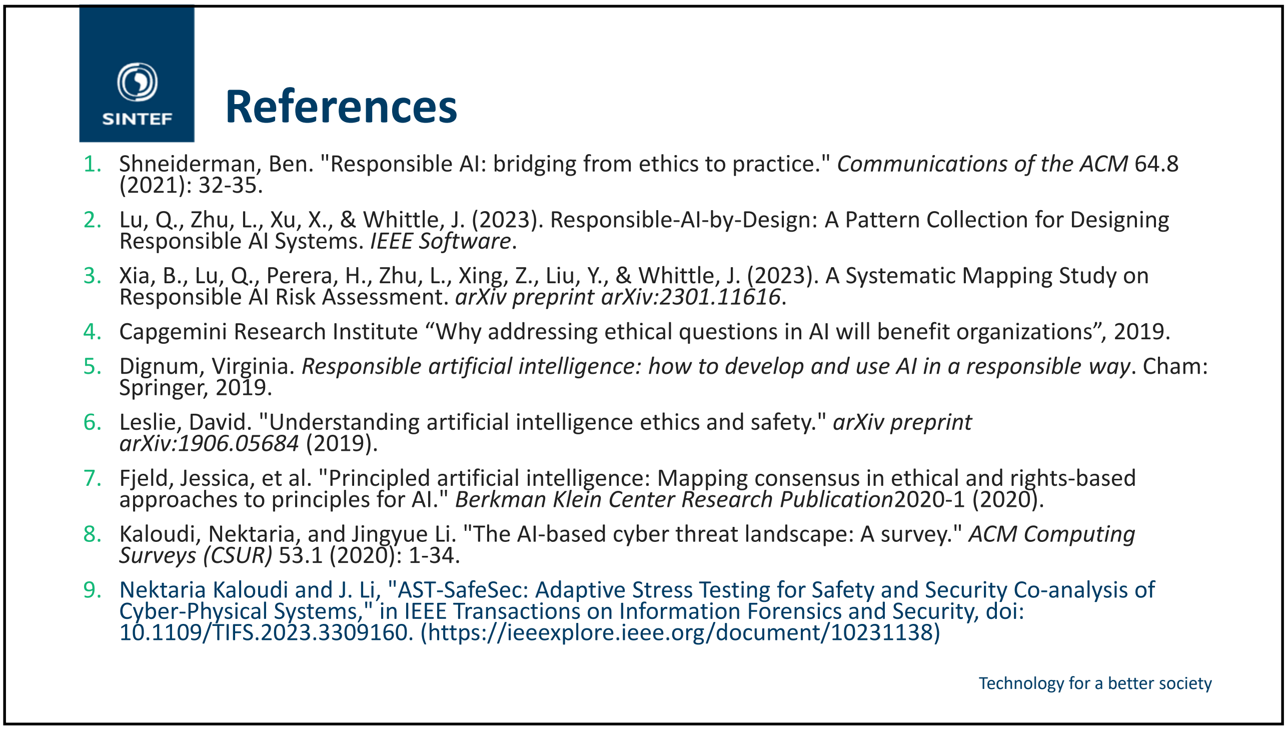
SINTEF

Conclusions

- **Trust the AI** is related to its ability to operate within certain constraints, and AI should know its constraints and provide warnings when it cannot be trusted
- Tensions and **trade-off analysis** between the various principles should be conducted
- Solutions can **address multiple Responsible AI principles** in a unified way or in-parallel way
- **Ethical risk assessment frameworks** tailored to ML system that consider the continuously learning capabilities
- **Translation of requirements into test cases** and continuously checked, including context awareness
- The need for a responsible and **human-centric approach** to ensure the development of ML systems will be aligned with Responsible AI principles

Technology for a better society

24



References

1. Shneiderman, Ben. "Responsible AI: bridging from ethics to practice." *Communications of the ACM* 64.8 (2021): 32-35.
2. Lu, Q., Zhu, L., Xu, X., & Whittle, J. (2023). Responsible-AI-by-Design: A Pattern Collection for Designing Responsible AI Systems. *IEEE Software*.
3. Xia, B., Lu, Q., Perera, H., Zhu, L., Xing, Z., Liu, Y., & Whittle, J. (2023). A Systematic Mapping Study on Responsible AI Risk Assessment. *arXiv preprint arXiv:2301.11616*.
4. Capgemini Research Institute “Why addressing ethical questions in AI will benefit organizations”, 2019.
5. Dignum, Virginia. *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Cham: Springer, 2019.
6. Leslie, David. "Understanding artificial intelligence ethics and safety." *arXiv preprint arXiv:1906.05684* (2019).
7. Fjeld, Jessica, et al. "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI." *Berkman Klein Center Research Publication*2020-1 (2020).
8. Kaloudi, Nektaria, and Jingyue Li. "The AI-based cyber threat landscape: A survey." *ACM Computing Surveys (CSUR)* 53.1 (2020): 1-34.
9. Nektaria Kaloudi and J. Li, "AST-SafeSec: Adaptive Stress Testing for Safety and Security Co-analysis of Cyber-Physical Systems," in IEEE Transactions on Information Forensics and Security, doi: 10.1109/TIFS.2023.3309160. (https://ieeexplore.ieee.org/document/10231138)

Technology for a better society

25

Technology for a better society

26