ASAF
27.1.2023

Itsekehittyvät järjestelmät –
Ajatuksia tekoälyyn liittyvistä turvallisuusvaatimuksista

Self-evolving systems
AI, Machine Learning

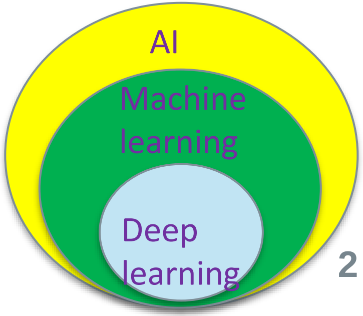
Timo Malm
VTT Technical Research Centre of Finland Ltd.

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

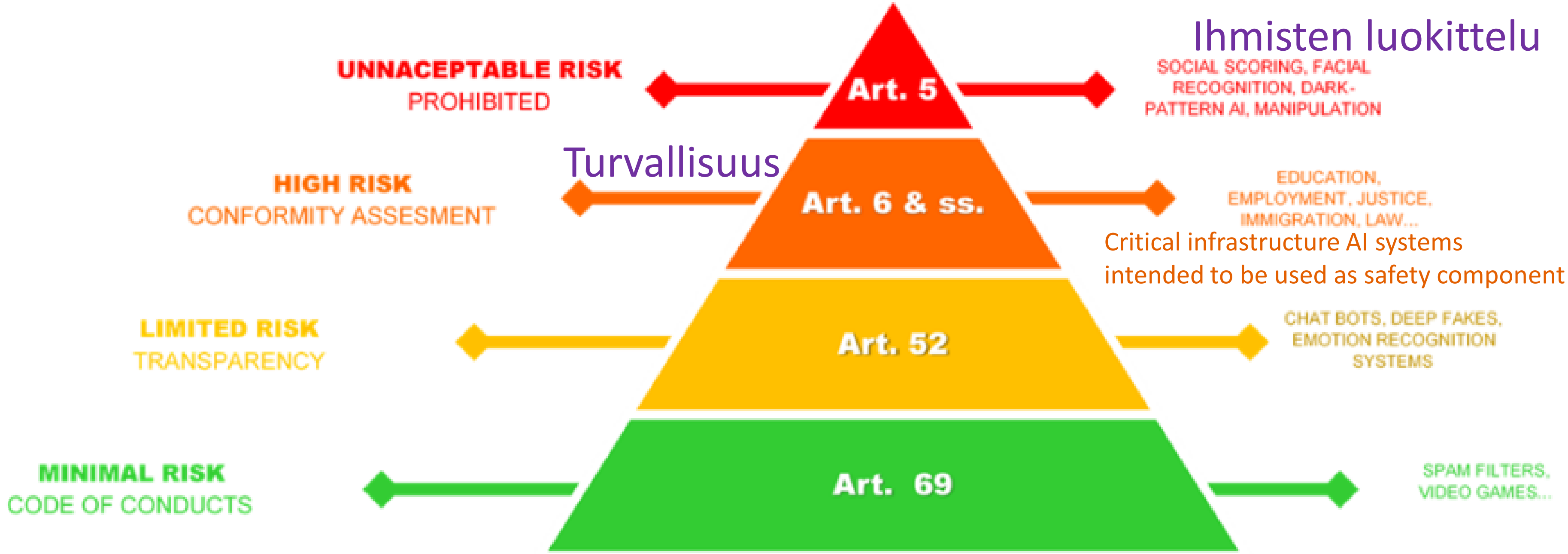
https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

- ‘**artificial intelligence system**’ (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with;
- Very generally speaking, in control systems **Machine Learning** = learning from examples.
- *“**Artificial intelligence** is that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment.”* – Nils J. Nilsson
- John McCarthy is one of the "**founding fathers**" of artificial intelligence, together with Alan Turing, Marvin Minsky, Allen Newell, and Herbert A. Simon. McCarthy, Minsky, Nathaniel Rochester and Claude E. Shannon coined the term "artificial intelligence" in a proposal that they wrote for the famous Dartmouth conference in Summer 1956. This conference started AI as a field. [Wikipedia]

AI > Self-evolving systems, machine learning > Deep learning
Self-evolving behaviour means that machine operation can change within defined limits.
The machine system learns to improve its performance during use or before the use.
Machine learning is a branch of statistics and computer science, which studies algorithms and architectures that learn from observed facts. [Wikipedia]



The AI Act’s hierarchy of risks



Ihmisten luokittelu

UNNACEPTABLE RISK
PROHIBITED

Art. 5

SOCIAL SCORING, FACIAL RECOGNITION, DARK-PATTERN AI, MANIPULATION

Turvallisuus

HIGH RISK
CONFORMITY ASSESMENT

Art. 6 & ss.

EDUCATION, EMPLOYMENT, JUSTICE, IMMIGRATION, LAW...

Critical infrastructure AI systems intended to be used as safety component

LIMITED RISK
TRANSPARENCY

Art. 52

CHAT BOTS, DEEP FAKES, EMOTION RECOGNITION SYSTEMS

MINIMAL RISK
CODE OF CONDUCTS

Art. 69

SPAM FILTERS, VIDEO GAMES...

Ref:Artificial Intelligence Act: What Is the European Approach for AI?
By Eve Gaumond Friday, June 4, 2021, 11:50 AM

Definitions the new Machinery proposal

European Comission– Parliament – Council (proposal)

Three slightly different proposals of legislation.

Definitions

(3) ‘safety component’ means a component of physical, digital or mixed nature component, including software, of products subject to this Regulation which serves to fulfil a safety function and which is independently placed on the market, the failure or malfunction of which endangers the safety of persons but which is not necessary in order for the machinery products subject to this Regulation to function or may be substituted by normal components in order for the products subject to this Regulation to function;

Kolme eri versiota EU:n koneasetuksesta.

Julkaistaneen vuoden parin sisällä

Machinery regulation Annex III section 1.2.1

- Control systems of machinery or related product with **fully or partially self-evolving behaviour** or logic that is designed to operate with varying levels of autonomy shall be designed and constructed in such a way that:
  - (a) they shall not cause the machinery or related product to perform actions **beyond its defined task** and movement space;

Pitää määritellä tarkasti AI:n (itsekehittyvä järjestelmä) rajat, mitä ei saa ylittää. Esim. AI ohjattu mobiilirobotti ei saa ylittää sille määriteltyä aluetta, nopeutta tai generoida uusia määrittelemättömiä tehtäviä.

Machinery regulation Annex III

- B. GENERAL PRINCIPLES: The risk assessment and risk reduction shall include hazards that may be generated during the lifecycle of the machinery or related product that are foreseeable at the time of placing of the machinery or related product on the market as an intended evolution of its fully or partially self-evolving behaviour or logic as a result of the machinery or related product designed to operate with varying levels of autonomy.

- Ergonomy (1.1.6): adapting the human-machine interface to the foreseeable characteristics of the operators, including with respect to a machinery or related product with intended fully or partially self-evolving behaviour or logic that is designed to operate with varying levels of autonomy;

Riskin arvioinnissa otetaan huomioon järjestelmän oppiminen.

ANNEX I
HIGH-RISK CATEGORIES OF MACHINERY OR RELATED PRODUCTS TO WHICH ONE OF THE PROCEDURES REFERRED TO IN ARTICLE 21 (2) AND (2A) SHALL BE APPLIED

Part A Categories of machinery or related products to which the procedures referred to in Article 21 (2) [*Conformity assessment*] shall be applied

Kaupalliset

- 24. Safety components with **fully or partially self-evolving behaviour** using machine learning approaches or logic Systems ensuring safety functions.

Omat yms.

- 25. Machinery embedding Systems with fully or partially self-evolving behaviour using machine learning approaches ensuring safety functions that have not been placed independently on the market, in respect only to those systems.

Tyyppitarkastukset

Conformity assessment:

(a) EU type-examination procedure (module B) provided for set out in Annex VII, followed by conformity to type based on internal production control (module C) set out in Annex VIII;

(b) Conformity based on full quality assurance (module H) set out in Annex IX;.

(c) Conformity based on unit verification (module G) set out in Annex IXa. *Conformity based on unit verification is the conformity assessment procedure whereby the manufacturer fulfils the obligations laid down in points 2, 3 and 5, and ensures and declares on his or her sole responsibility that the machinery or related product, which has been subject to the provisions of point 4, is in conformity with the essential health and safety requirements set out in Annex III. The manufacturer shall establish the technical documentation and make it available to the notified body…A notified body chosen by the manufacturer shall carry out appropriate examinations and tests…*

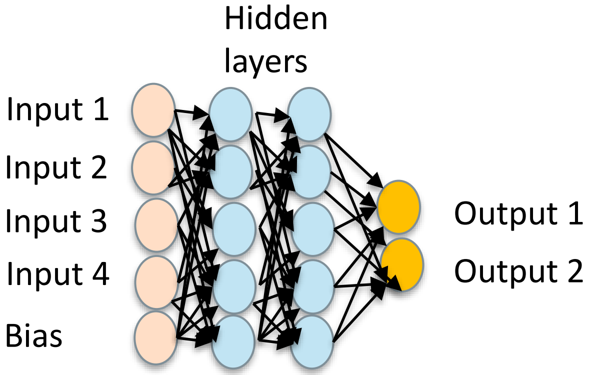
ISO/TR 22100-5:2021. Safety of machinery — Relationship with ISO 12100 — Part 5: Implications of artificial intelligence machine learning

- 3.1 **artificial intelligence**, AI: branch of science devoted to developing data processing systems that perform functions normally associated with human intelligence, such as reasoning, learning, and self-improvement
  päättelyä, oppimista, itsekehitystä
- 3.2 **machine learning**: process using algorithms rather than procedural coding that enables learning from existing data in order to predict future outcomes datasta oppimista
- Enterprises in the machinery sector are constantly developing AI solutions for different application processes, such as:
- a) quality control;
- b) process optimization;
- c) condition/failure monitoring;
- d) predictive maintenance.

- Examples
- 4.2.1.2 Optimization of herbicide spraying machine
- 4.2.1.3 Optimization of a laser cutting machine
- Voice-controlled laser with AI (by TRUMPF).
- An example with safety implications is an automated guided vehicle (**AGV**) that self-optimizes its navigation via an AI application.
- **Conclusions**: The risk(s) introduced by AI in machinery applications can be completely **addressed by the methodology for risk assessment and risk reduction as prescribed in ISO 12100** where risks of the AI are addressed according to the **intended use and use limits** (predetermined boundaries) specified by the machine manufacturer.
  Voidaan soveltaa riskin arviointia (ISO 12100), kun otetaan huomioon tarkoitettu käyttö ja käyttörajat

Neural networks (Neuroverkot)

1. Rakennetaan malli
2. Valitaan data opetusta varten
3. Opetetaan verkkoa (kertoimet/kaavat neuroneille) datalla
4. Testataan ja parannetaan verkkoa toisella datalla
5. Käytetään verkkoa ja mahdollisesti opetetaan verkkoa samalla



Multilayer Perceptron Neural Network

Structure of AI and risks related to AI

Steps to design an AI system

1. Identify the problem.
2. Prepare the data.
3. Choose the algorithms.
4. Train the algorithms.
5. Test the model
6. Run on a selected platform.

Gaurav Chauhan

AI Safety

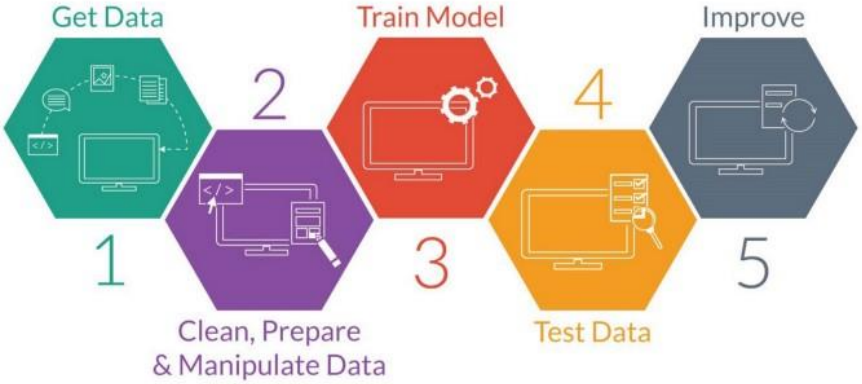
AI Safety is collective termed ethics that we should follow so as to avoid problem of accidents in machine learning systems, unintended and harmful behavior that may emerge from poor design of real-world AI systems. Adversarial attack are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake.

Examples of risks

Unbalanced data?
Fairness? Relevant?

Model type and accuracy?

New biased data?



Biased data?
Unilateral data?

Structured/unstructured data?
Manipulating principles?
Transparency?

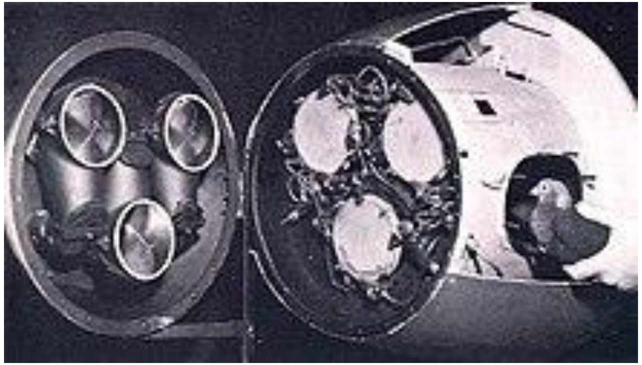
Ref: A simple way to explain how to build an AI system
Roger Chua. May 30, 2019 ·

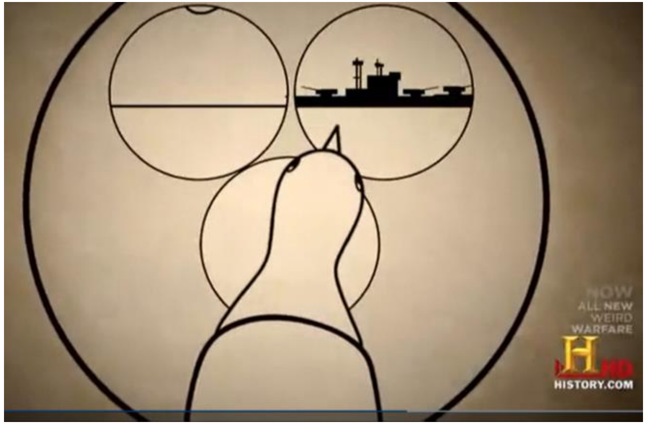
Ei ihan tekoälyä – Skinner 1944 – Pigeon-project

”bioäly”

- Toisessa maailmansodassa pulut opetettiin ohjaamaan ohjusta.
- Ohjuksessa 3 pulua, jotka nokkivat kuvaruudun maalia, johon ohjus kulkeutui. Projekti päättyi ennen laukaisuja.
- Tarvittiin kolminkertainen järjestelmä riittävän luotettavuuden saamiseksi ja toisaalta palkitsemisviiveisiin.

Muistuttaa tekoälyä: opetusprosessi, palkitseminen, ohjaus, takaisinkytkentä, redundanssia…
Ehkä tästä voisi oppia jotain?

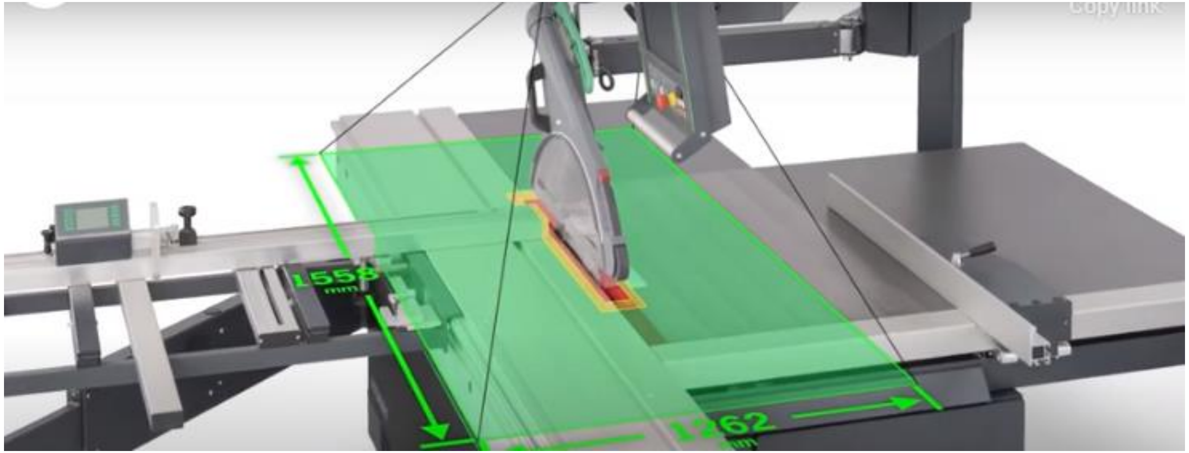




[Wikipedia, History.com/Military.com video]

AI is becoming more common
- There are already AI devices for safety purposes

Sophia (Hanson Robotics) – First robot/AI to get citizenship 2017 (Saudi Arabia).



AI-based safety device

Altendorf Hand Guard - our safety system for sliding table saws

Concrete Problems in AI Safety

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, Dan Mané

- Avoiding Negative Side Effects: achieving goal causes side effects
- Avoiding Reward Hacking: e.g. rewarding disables senses
- Scalable Oversight: e.g. missing object values cause work overload
- Safe Exploration: e.g. dangerous exploratory moves
- Robustness to Distributional Shift: e.g. strategi shifts when environment changes

- Wrong objective function (“avoiding side effects" and “avoiding reward hacking"),
- Objective function that is too expensive to evaluate frequently (“scalable supervision") or
- Undesirable behavior during the learning process (“safe exploration“ and “distributional shift").

By Google etc.

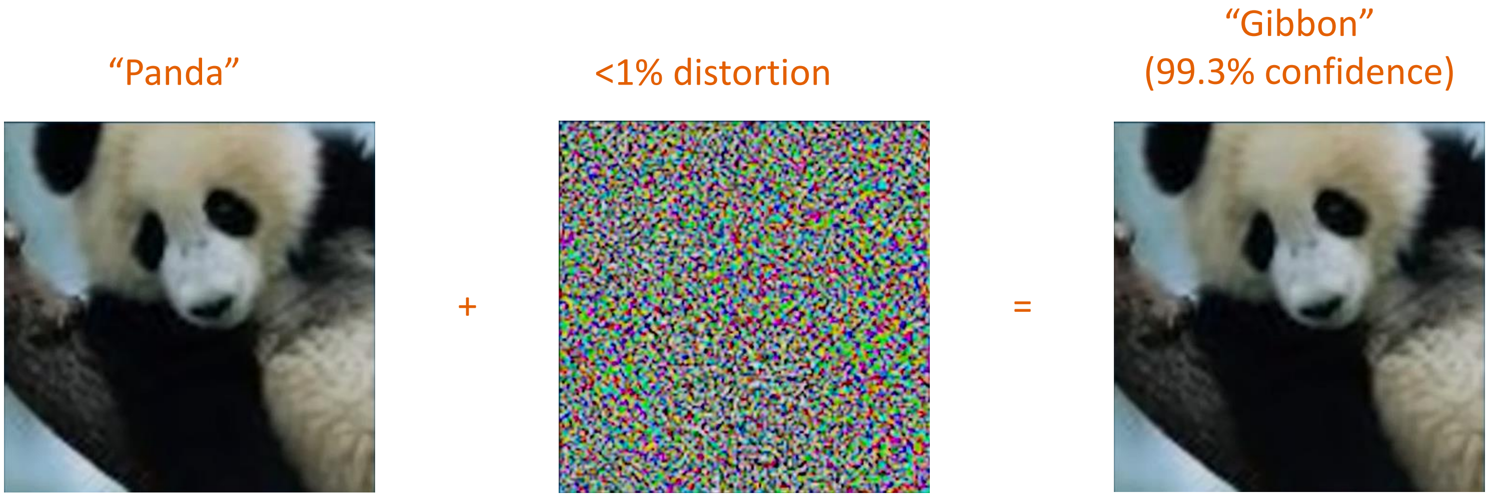
Gaurav Chauhan: AI Safety

AI Safety is collective termed ethics that we should follow so as to avoid problem of accidents in machine learning systems, unintended and harmful behavior that may emerge from poor design of real-world AI systems.

Adversarial attack are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake.

So how many pixels we need to change in a picture to fool our neural network?
Unfortunately, the answer is one.
To find a weak point of the neural network may require access to confidence values, but random values…

Security issues: Adversarial attacks



- Distortions can be crafted to produce the desired erroneous outcome

Jos palkitaan häiriöllisistä kuvista (esim. otettu kaukaa) voi tulos vääristyä

Ref. Eetu Heikkilä

Example from https://www.darpa.mil/about-us/darpa-perspective-on-ai

“Intelligence-related problems”

Reward hacking

- AI system aims to fulfil a pre-defined goal, and is “rewarded” when achieving it
- How to ensure that there are no loopholes in the system that allow unintended rewards?
- Example: A floor-washing robot is rewarded when no stains are detected on the floor. What if the robot just disables its vision to not see the stains, or covers the floor with other material to hide the stains?

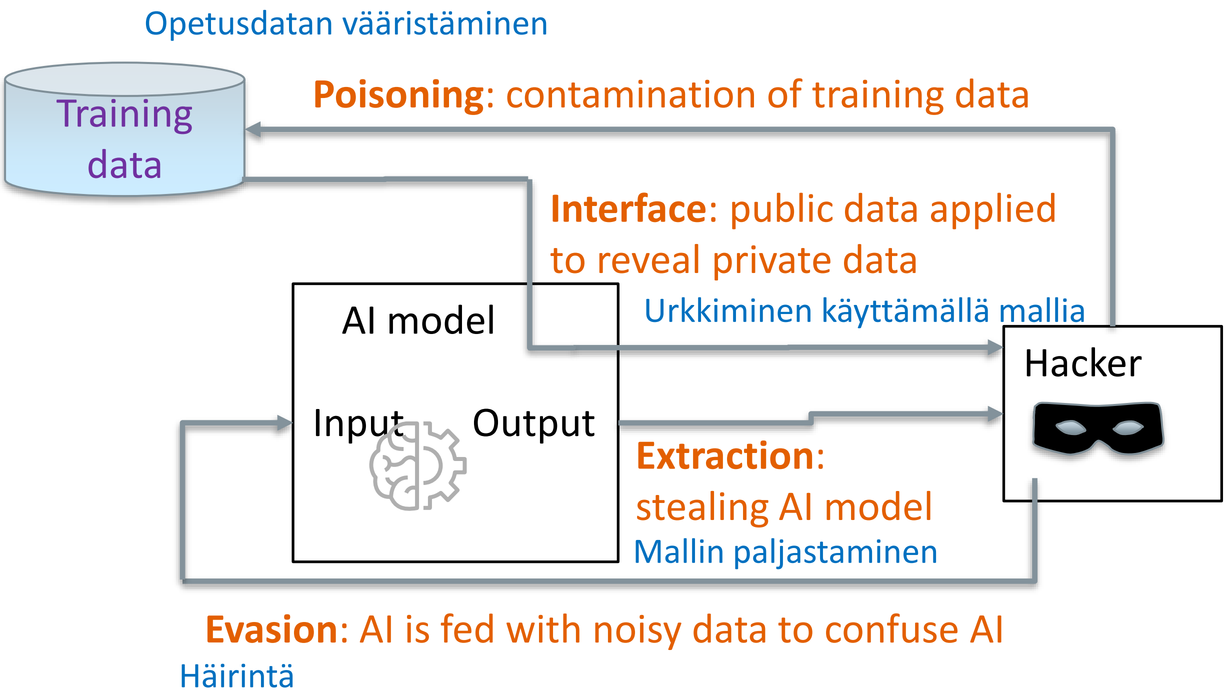
Palkinto saadaan, kun ei näy tahroja

Supistamalla näkökenttää tahroja ei näy

Ref. Eetu Heikkilä

Cybersecurity threats to AI

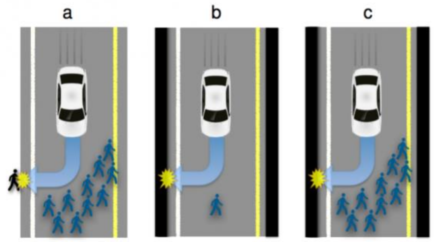
These attack patterns aim to collected proprietary information of target's processes or impact into AI based decision making within the factory processes.


Opetusdatan vääristäminen
Training data
Poisoning: contamination of training data
Interface: public data applied to reveal private data
Urkkiminen käyttämällä mallia
AI model
Input
Output
Hacker
Extraction: stealing AI model
Mallin paljastaminen
Evasion: AI is fed with noisy data to confuse AI
Häirintä

Ref: Some aspects of preventing AI manipulation. CF#1 Webinar, April 28th, 2021
Seppo Heikura, Board member, Senior Advisor, Houston Analytics

Ethics & safety: more than individual accident scenarios!

- A huge number of scenarios would need to be developed
  - Which are the ones to consider?
  - Is it acceptable to end up in certain scenarios in the first place?
- The ethics discussion shouldn't get stuck around individual, often trivial cases



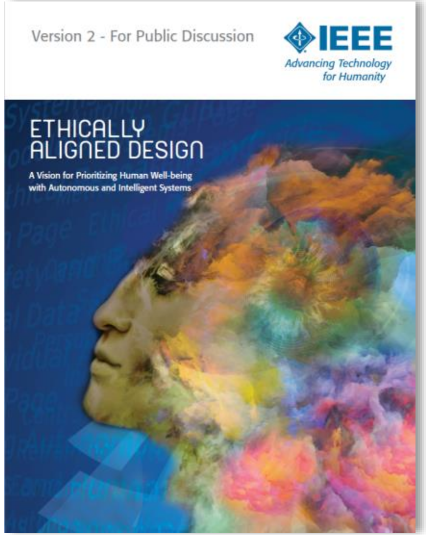
Onko oikein laskea kuolemantapausten lukumäärää? Pitäisikö palkita kävelytien käytöstä?
Pitäisikö suojella enemmän matkustamon väkeä vai jalankulkijoita,
koska tieto jalankulkijoista on epävarmaa?

Ref. Eetu Heikkilä

Standardization is only starting to develop

VTT

- ISO/IEC JTC 1/SC 42 - Artificial intelligence
- IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems
- SAE CRB 1-2016 (SAE CRB1-2016): Managing the Development of Artificial Intelligence Software
- **So far, general guidelines, rather than prescriptive standards**


Version 2 - For Public Discussion
IEEE
Advancing Technology for Humanity
ETHICALLY ALIGNED DESIGN
A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems

- ISO/IEC TS 4213 Assessment of machine learning classification performance
- ISO/IEC CD 5259 Data quality for analytics and machine learning (ML)
- ISO/IEC DTR 5469 Functional safety and AI systems
- ISO/IEC AWI TS 8200 Controllability of automated artificial intelligence systems
- ISO/IEC AWI TS 5471 Quality evaluation guidelines for AI systems
- ISO/IEC FDIS 23894 Guidance on risk management
- ISO/IEC DIS 5338 AI system life cycle processes
- ISO/IEC AWI TS 6254 Objectives and approaches for explainability of ML models and AI systems

Tekoälyn validointi turvallisuuteen liittyvissä ohjauksissa ei ole vielä näköpiirissä

P7000 – Model Process for Addressing Ethical Concerns During System Design
P7001 – Transparency of Autonomous Systems
P7002 7002 – Data Privacy Process
P7003- Algorithmic Bias Considerations
P7004 Standard for Child and Student Data Governance
P7005 Standard for Transparent Employer Data Governance
P7006 Standard for Personal Data Artificial Intelligence (AI) Agent
P7007 Ontological Standard for Ethically Driven Robotics and Automation Systemsma
P7008 Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems
P7009 Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems
P7010 Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems

Ref. Eetu Heikkilä, Saeed Bakhshi Germi

Some AI failures

- Microsoft Tay turns to racist and sexist. The goal was to build a slang-filled chatbot that would raise machine-human conversation quality to a new level.
- Amazon recruiting tool was not gender-neutral
- Inverness Caledonian Thistle F.C. Ball Tracking System failed to track ball (bald head confusion).
- Uber Self Driving Car Fatality, Tesla cars crash due to autopilot feature
- Face ID Hacked Using a 3D Printed Mask
- False facial recognition match leads to Black man’s arrest
- Fail: IBM’s “Watson for Oncology” Cancelled After $62 million and Unsafe Treatment Recommendations (cancer)

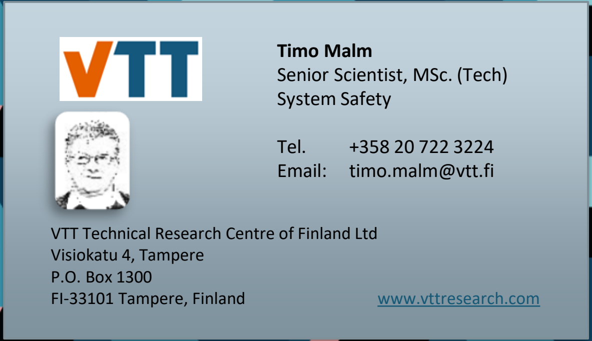
Turvallisuuteen liittyvien ohjausjärjestelmien kehitys

Tekniikka	Standardeja	Aika
Pakkotoimiset releet		1970
Turvareleet, muuntajat, elektroniikan piirirakenteet, oskillaatio vikojen havaitsemiseen		1980
Piirirakenteiden validointi, Ohjelmoitavat turvalogiikat, elinkaarimalli	EN 954-1:1996, IEC 61508:1998	1990
Turvaväylät, sarjamuotoinen tieto	EN 50159, IEC 61508-2 IEC 61784-3:2007	2000
Tekoäly turvapiireissä (?) Hahmontunnistus, paikantaminen, navigointi, ohjaus…	Validointi? Opetusdata, tekoälyn tyyppi, malli, testidata	?

Conclusions

- It is very promising that AI or self-evolving systems can solve many current problems.
- It is currently very difficult to validate AI. It may make failures that ordinary logic cannot. It can be difficult for a person to understand the failures that AI can do.
- “AI effect” – “AI brings a new technology into the common fold, people become accustomed to this technology, it stops being considered AI, and newer technology emerges”

Thank you for your attention!



Timo Malm
Senior Scientist, MSc. (Tech)
System Safety

Tel. +358 20 722 3224
Email: timo.malm@vtt.fi

VTT Technical Research Centre of Finland Ltd
Visiokatu 4, Tampere
P.O. Box 1300
FI-33101 Tampere, Finland

www.vttresearch.com