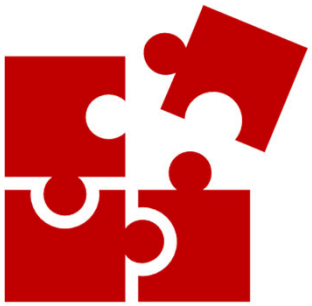
Sist oppdatert 6.9.2023

TTK 2 – Design og analyse for funksjonell sikkerhet 2023

Seminar 3 – 06.09.2023

MARY ANN LUNDTEIGEN
TEKNISK KYBERNETIKK NTNU



3. AI og sikkerhetssystemer

Med hjelp av PhD student Rialda Spahic

Motivasjon

I dag **tillates** normalt **IKKE** kunstig intelligens industrielle sikkerhetskritiske systemer. Samtidig bør kunstig intelligens (KI/AI) ha **potensialet** til å bidra til **økt sikkerhet**. Men **uforutsigbarheten** i oppførselen og **usikkerheten om evnene** til KI/AI systemer må reduseres og funksjonsdyktigheten må kunne dokumenteres.

Vi må derfor finne måter å bruke kunstig intelligens på en trygg måte også for sikkerhetskritiske systemer.

Litteratur

Pensum:

- Slides

Støttelitteratur:

- IEC TR 5469 (utkast)
- AI + Safety (DNV paper)
- Annet lenket opp (litt på EU regelverksforslag)

AI og AI metoder

Hva er kunstig intelligens (AI)

Figur: Arcinsightpartner

AI is the study and development of computer systems that can **copy intelligent human behavior**.
The Oxford Advanced Learner's Dictionary.

«KI er nemlig ikke en teknologi i seg selv. Det er heller en kategori – en samlebetegnelse – som rommer ulike teknikker og teknologier som produserer noe vi oppfatter som **intelligent oppførsel fra en datamaskin**.» KI kan ikke tenke selv, men gjøre ting vi ikke forstår.
Inga Strümke (NTNU)

Ilustrasjoner lagt til figuren som er lånt fra Inga Strümke

Forsterket, dyp, veiledet og ikke veiledet ML

Maskinlæring (ML): **Algoritmer** som **lager modeller** som gir svar på et **bestemt og angitt problem** ved bruk av **data.**

- **<u>Forsterket (reinforced) læring:</u>**
  Bruker belønning for å utviklet en modell
- **<u>Veiledet (supervised) læring:</u>**
  Bruker fasit til å lage modell
- **<u>Ikke-veiledet (unsupervised) læring:</u>**
  Lager modell basert på mønstre og sammenhenger uten å vite noe om fasiten
- **<u>Dyp læring:</u>**
  Lager modell som består av underliggende forklaringsmodeller som ikke er synlig utad. Dyp læring kan være forsterket, veiledet og ikke-veiledet læring basert på store datamengder .

Maskinlæring og målet til modellene

Maskinlæring (ML): **Algoritmer** som **lager modeller** som gir svar på et **bestemt og angitt problem** ved bruk av **data**

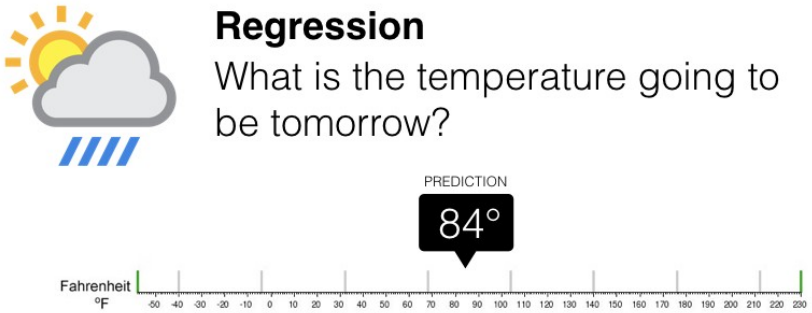
Modellene er enten utviklet for å utføre:

- **klassifisering** (resultatet er å skille ut, kategorisere noe)

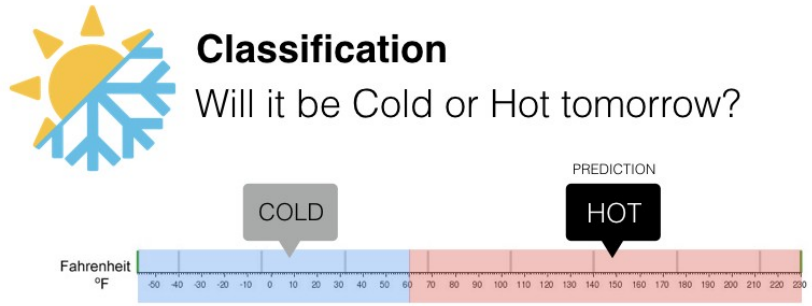
  eller

- **regresjon** (gir et resultat basert på styrker og sammenhenger i data)

Bruksområder: Regresjon og klassifisering



Regresjon predikterer en **bestemt verdi på et fenomen eller størrelse** vi ønsker å bestemme.

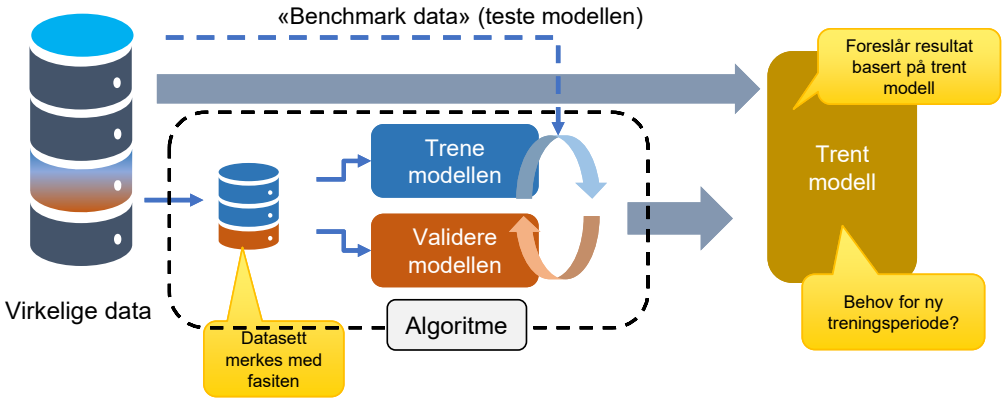


Klassifisering **plasserer** et fenomen eller størrelse **i en av flere angitte klasser/kategorier**

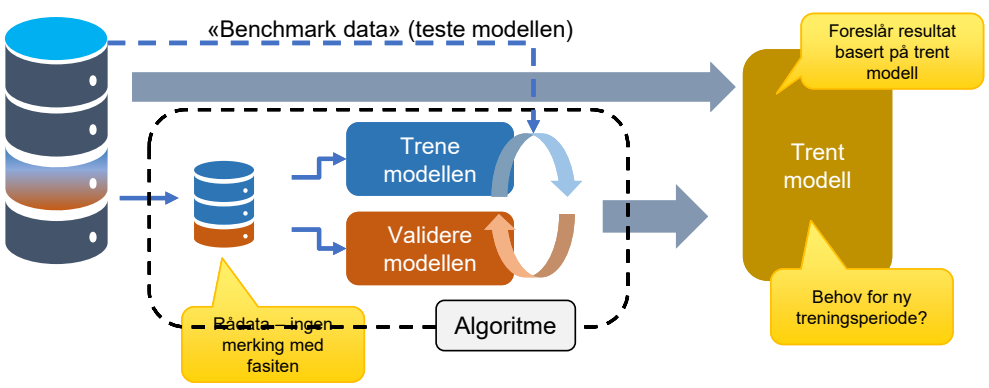
Kilde: https://www.springboard.com/blog/data-science/regression-vs-classification/

Utvikling av ML modeller

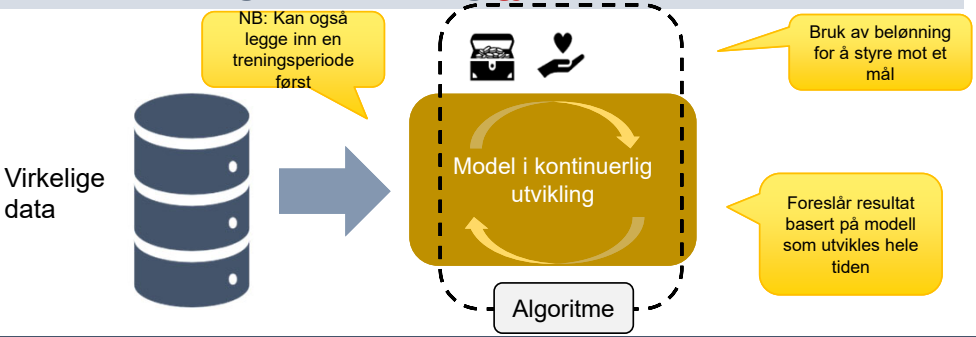
Veiledet læring - Læring *før* bruk



Ikke-veiledet læring – Læring *før* bruk



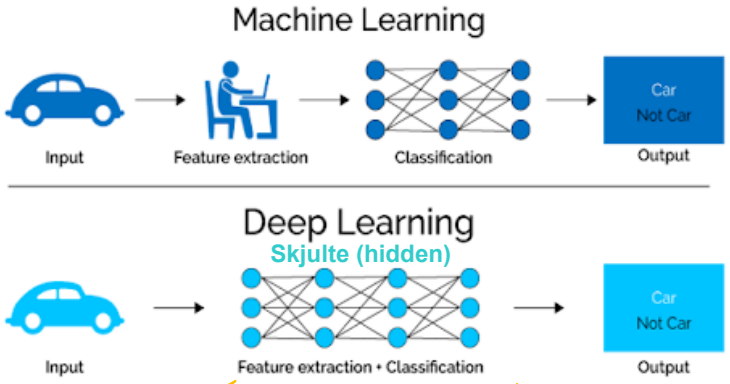
Forsterket læring - Kun læring *gjennom* bruk



Algoritme: Måten modellen utvikles på

Dyp læring (deep learning)

Dyp læring (deep learning): En variant av maskinlæring som baserer seg på algoritmer som imiterer læringsprosessen i hjernen (nevroner og forbindelser) og som i stor grad bruker «big data» (store datamengder)



Kan brukes for veiledet, ikke veiledet, og forsterket læring

Kan selv klare å finne underliggende mønstre og sammenhenger

- To hovedkategorier av ANN: Convolutional neural networks (CNN)
- Recurrent natural networks (RNN)

Mange ANN metoder: self organizing maps, multilayer perceptrons,…

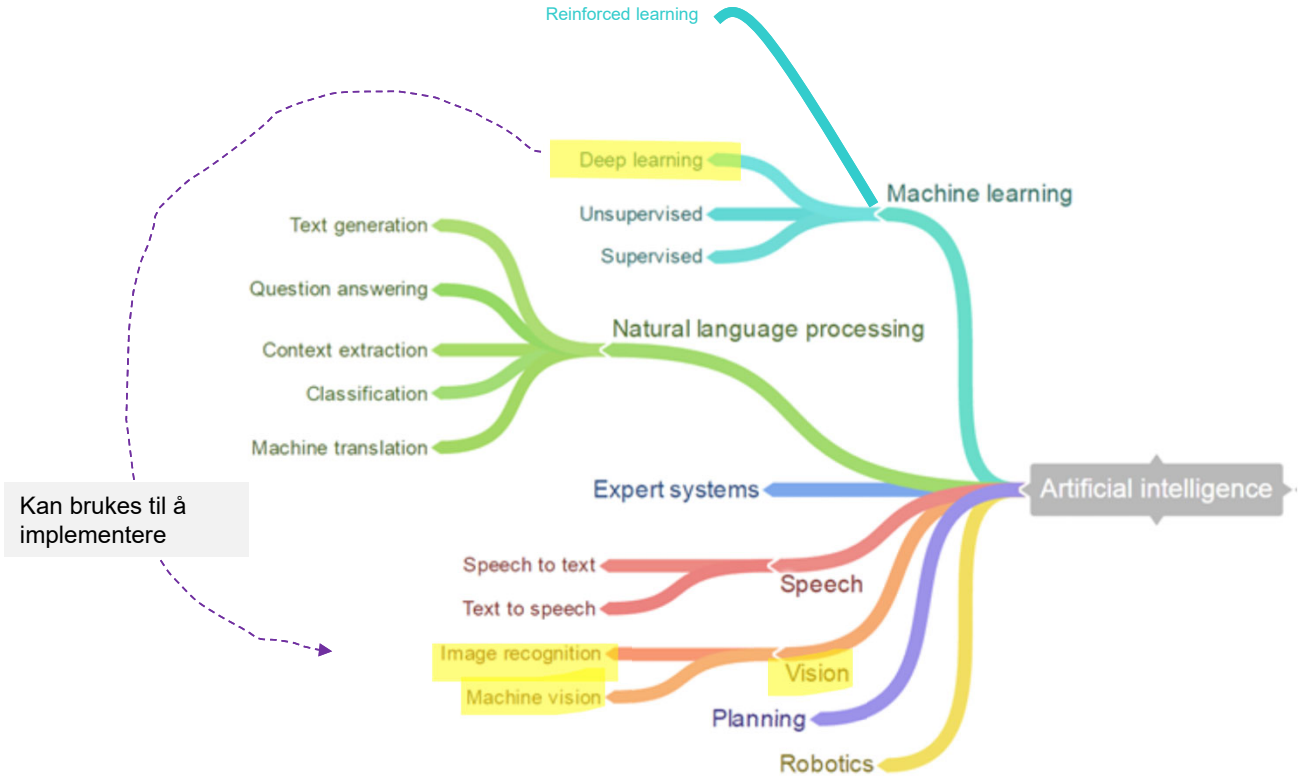
Bruksområder:

- Robotsyn (CNN)
- Talegjenkjenning (RNN)
- Taleanalyse (natural language procssing) (RNN)
- Medisinske analyser
- Klimaanalyser

Kilde: https://builtin.com/machine-learning/deep-learning

Kilde: https://en.wikipedia.org/wiki/Deep_learning

Klassifiseringen her er ikke egentlig helt presis

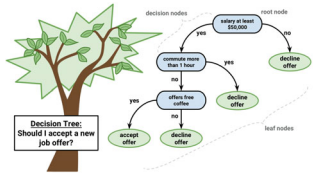


ML metoder /algoritmer (eksempler)

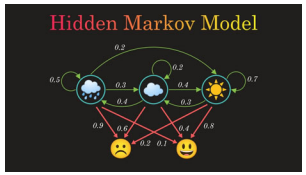
Metode/Algoritme	Type trening/læring: Supervised learning (trent/veiledet)	Unsupervised learning (utrent/ikke-veiledet)	Reinforced learning (forsterket)
Random forest (RF) [C,R]	X		
Decision tree (DT) [C]	X		
Hidden Markov model (HMM) [C]	**X**	X	
Deep learning: Artificial neural networks (ANN) [C,R]	**X**	**X**	X
Support vector regression (SVR) [R]	X		
Support vector machine (SVM) [**C**,R]	X		
Kernel regression [C]	X		
K-nearest neighbor (KNN) [C]	X		
K-Means Clustering [C]		X	
Principal Component Analysis (PCA) [R]		X	
Hierarchical Clustering [C]		X	
Markov Decision Process [O]	X		**X**
Bayesian belief networks (BBN) [C,O]	X		X
Bayesian regression [R,O]	X		?

C: Classification, R: Regression, O: Other (RL) – exploitation, exploration, policy learning, value learning,…

Decision tree (DT)

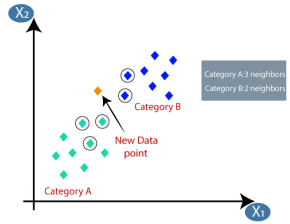


https://www.linkedin.com/pulse/decision-tree-gauransh-singh/



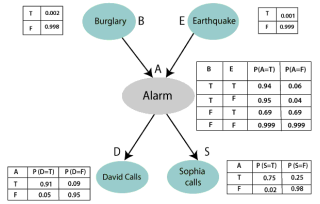
https://www.youtube.com/watch?v=RWkHJnFj5rY

K-nearest neighbour (KNN)

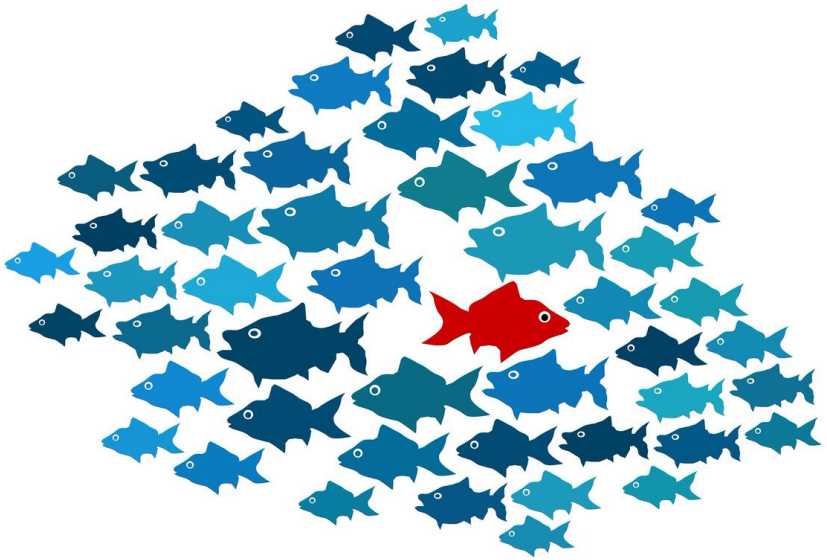


https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

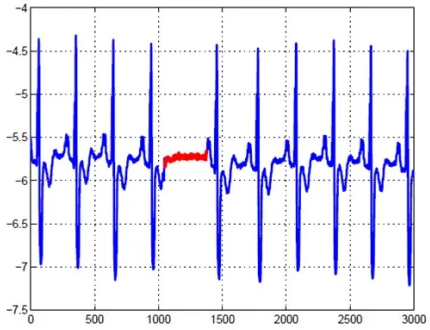
Bayesian belief networks

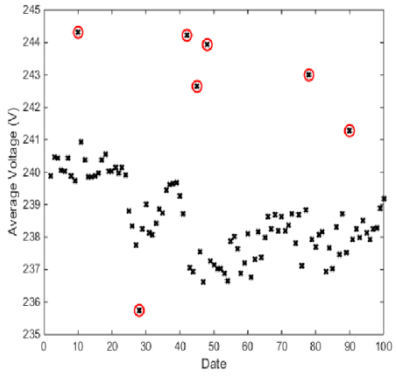


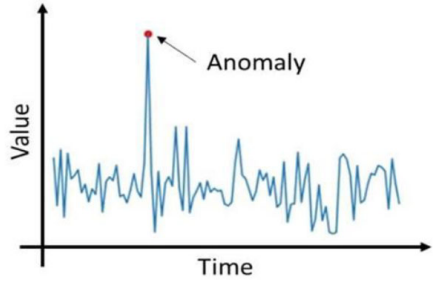
Eksempel på bruk av klassifisering: Finne anomaliteter



Datapunkt(er) utenfor forventet mønster, område,…

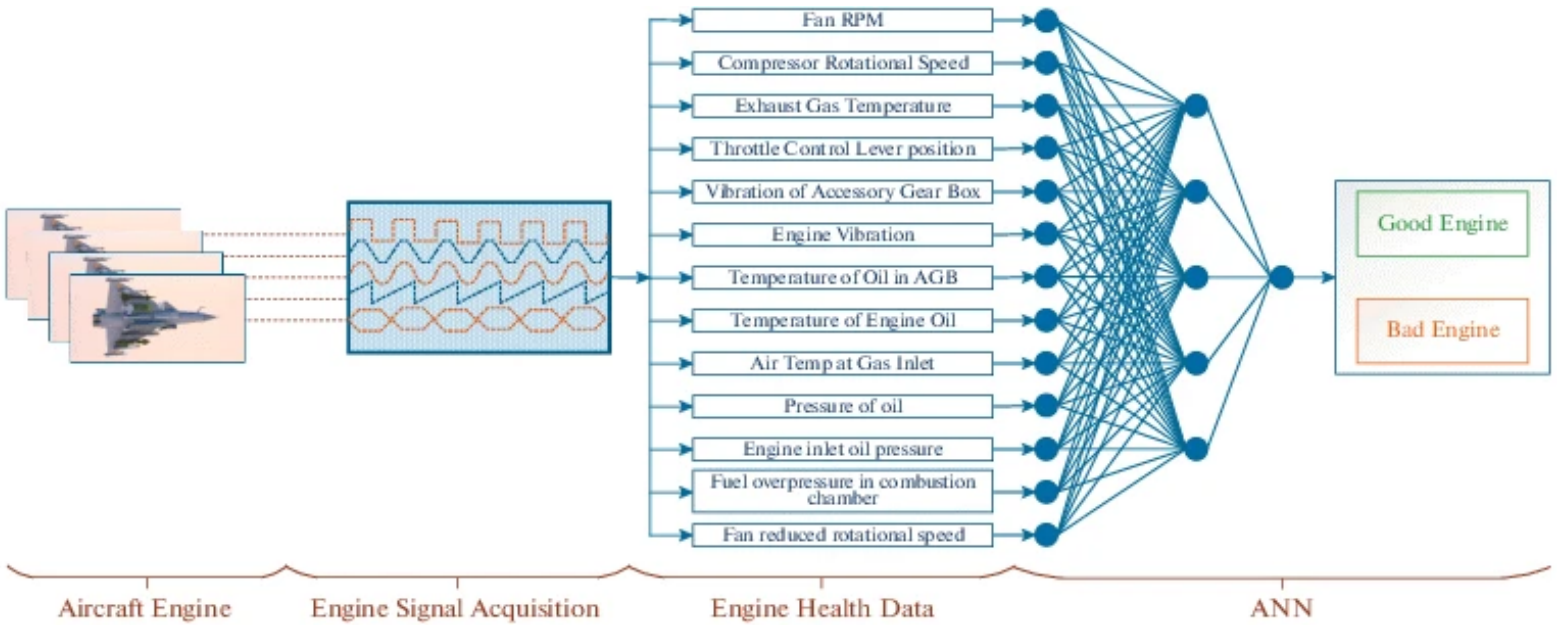






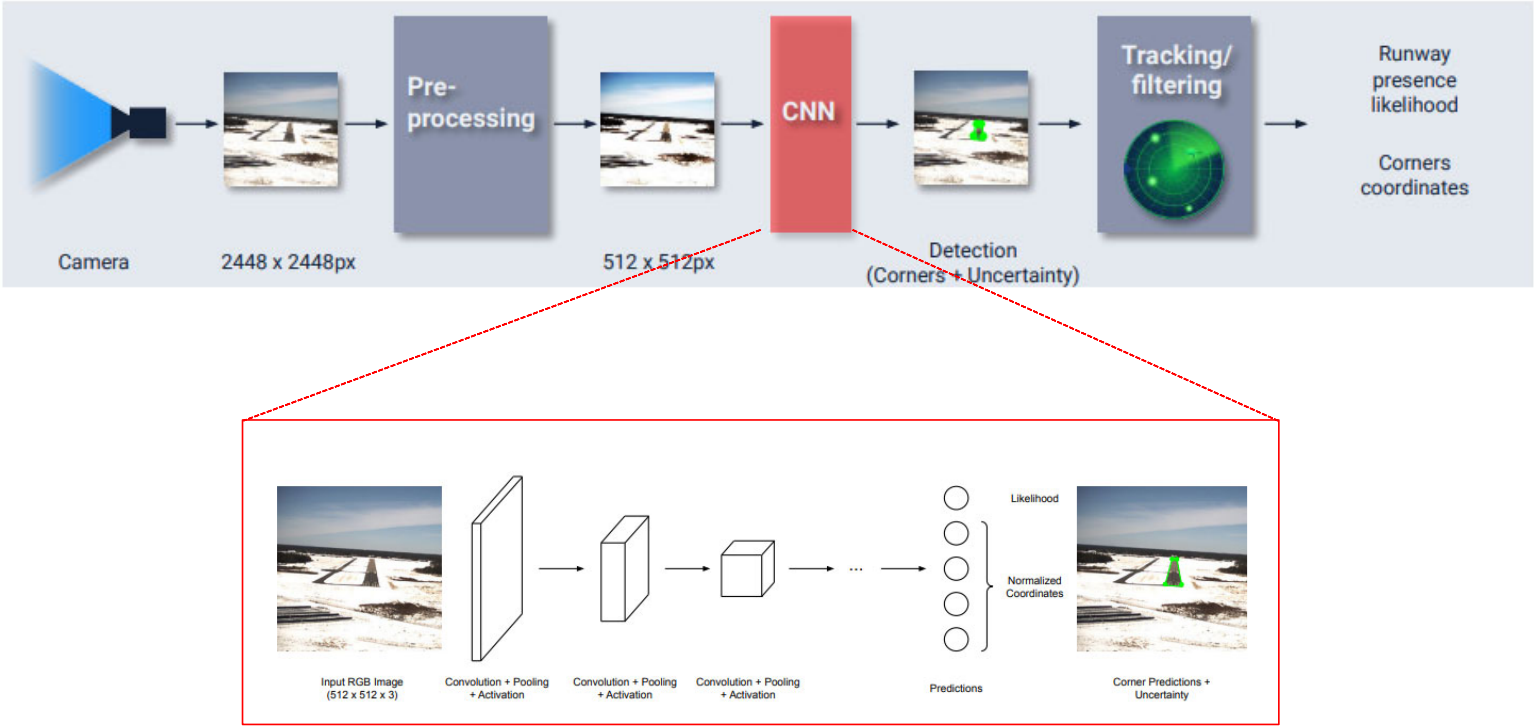
Kilde: https://thedatascientist.com/anomaly-detection-why-you-need-it/

Eksempel på bruk av dyp læring: Tilstandsovervåkning



Kilde: https://doi.org/10.3103/S1060992X21010094

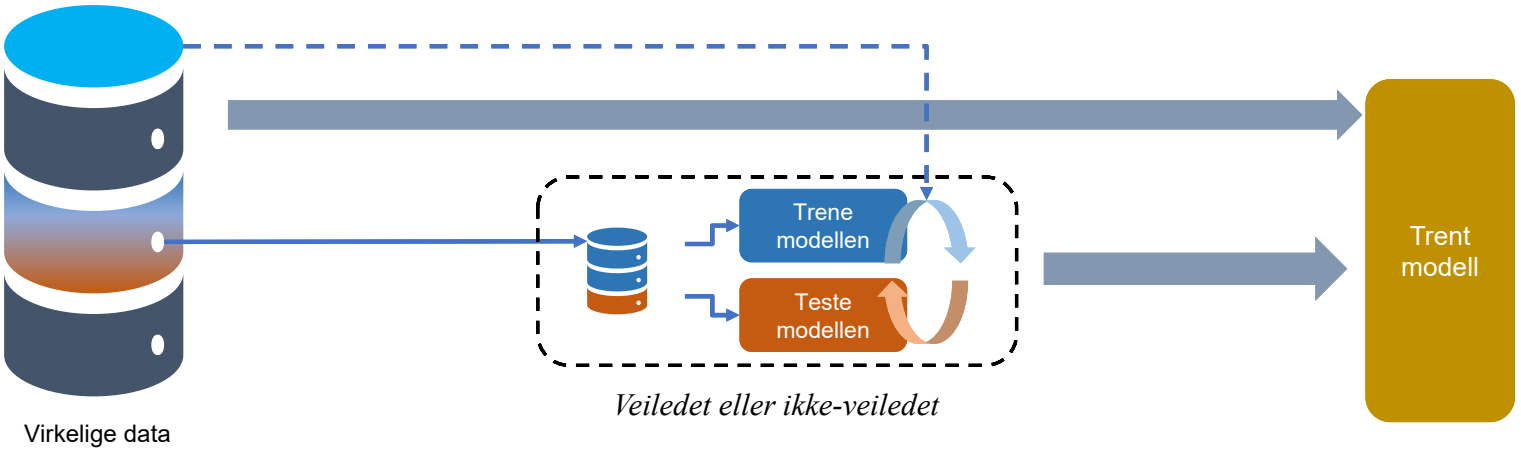
Eksempel på bruk av dyp læring: Visual landing guidance (VLG)



Kilde: https://www.easa.europa.eu/en/document-library/general-publications/concepts-design-assurance-neural-networks-codann

Maskinlæring: Grunner til at det kan gå galt…

Generelt for maskinlæring



1. **Historical bias** – misalignment between the world and values/objectives to the model
2. **Representation bias** – under-represented population
3. **Measurement bias** – bias that stem from using proxies of the desired quantities to collect

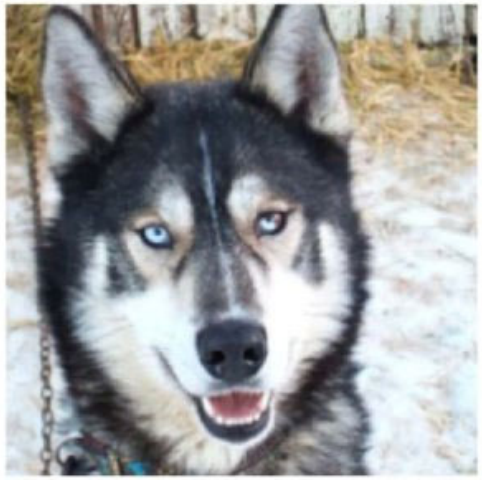
4. **Aggregation bias**: Arise during model constructing when distinct populations are inappropriately combined
5. **Evaluation bias;** Arise when benchmark and test data do not equally represent the various parts of the population. Not appropriate performance measure

6. **Deployment bias**: The system is used or interpreted the inappropriate way – leading to harmful, unlawful, or unethical decisions

Kilde: https://www.researchgate.net/publication/330726492_A_Framework_for_Understanding_Unintended_Consequences_of_Machine_Learning

Utfordring: Mest effektiv klassifisering = riktig kritierier brukt?

… eh, Nei. Her er et eksempel:



Hvordan skille Ulv eller Husky?

Dette kriteriet fant algoritmen til å være mest effektivt!

Algoritmen fant det mest effektivt å skille husky hund og ulv fra hverandre basert på om det var snø eller ikke i bakgrunnen (husky bilder hadde som regel snø i bakgrunnen)

Fungerte perfekt på alle bildene den hadde blitt trent på

Men ikke så robust da …

Kilde: https://www.researchgate.net/publication/305999024_Why_Should_I_Trust_You_Explaining_the_Predictions_of_Any_Classifier og https://arxiv.org/abs/1602.04938

Utfordring: Bruk av belønning – hva gir mest belønning?

Tay – en AI chatbot laget av Microsoft i 2016 med virtuell bruker på Twitter

Designet som:
Imitere språket til en typisk 19-årig amerikansk jente for å utføre interaksjon med brukere på twitter

Kilde: https://en.wikipedia.org/wiki/Tay_(chatbot)

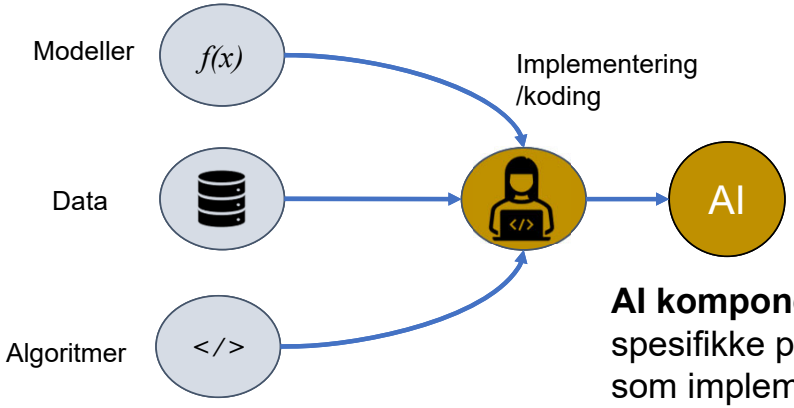
Ble til:
En aktiv rasist, provokatør og politisk ukorrekt «monster»

Mer interaksjon når man var provokatør enn hyggelig og «tannløs»

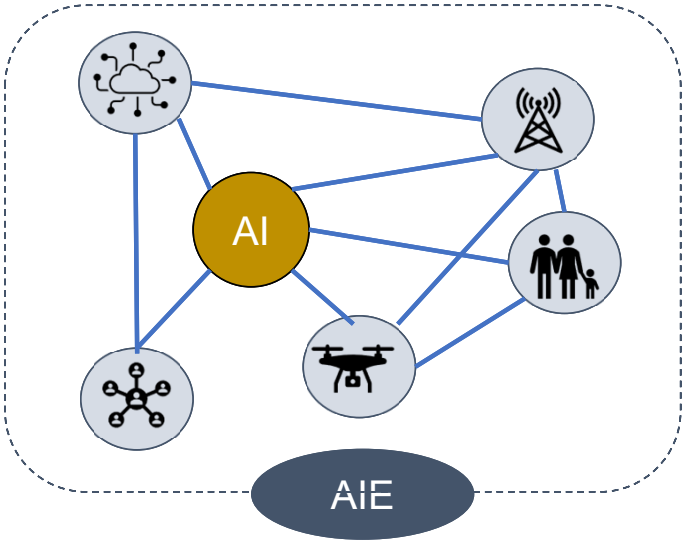
Påvirket av brukernes interesse for tema

Litt teoretisering om AI og krav som kommer i regelverk og retningslinjer

AI component og AI



AI komponent: Den spesifikke programkoden som implementerer AI modellen

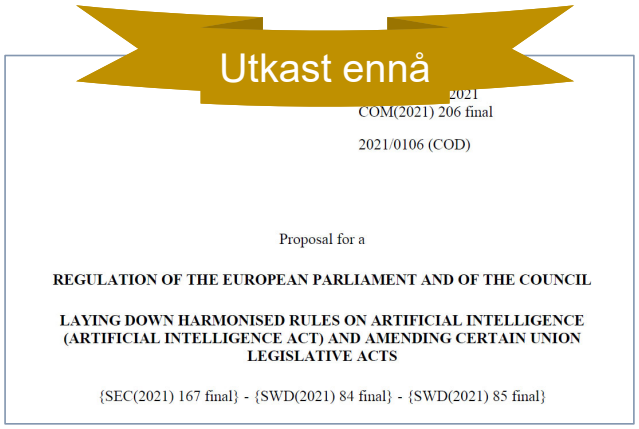


AI-enabled system (system med AI egenskaper): System der AI komponent inngår for beslutnings-taking, data prosessering, bildeanalyse, utføre autonome styringsfunksjoner, prediksjon av tilstand på utstyr,..

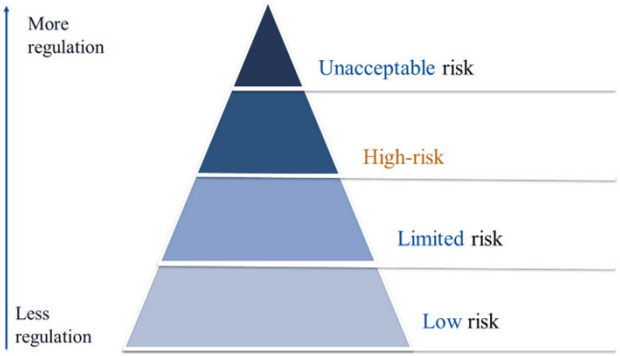
Kilde: DNV retningslinje **DNV RP 0771 (**under utgivelse).

Høyrisiko AI systemer (per EU)

- Omfatter systemer som skal ivareta **sikkerhet** eller systemer som kan påvirke **grunnleggende rettigheter**.
- Nytt lovforslag på høring i EU
- Fokus på å håndtere risikoforbundet med AI slik at man oppnår **tillitsverdig AI.**
- Legger opp til en **stegvis prosess** for å vise samsvar av AI komponent/AI enabled system m regelverk (m/ CE merke) før «produktet» blir tilgjengelig i markedet
  - Risikobasert tilnærming med krav om **risikostyring** gjennom hele livssyklusen
  - Krav til registrering i «EU database for «stand-alone high risk AI systems»

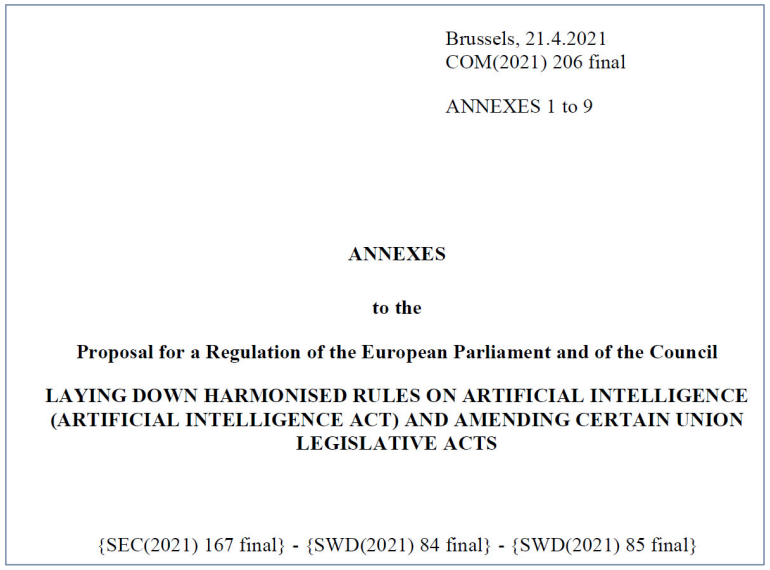


https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206



Eksempler på høyrisiko AI systemer (per EU)

- Biometrisk identifisering og kategorisering av personer
- Styring og operasjon av **kritisk infrastruktur** (veitrafikk, vann, gass, oppvarming, elektrisitet)
- Utvelgelse til opptak studier og vurdering av studenter
- Ansettelse og opprykk
- Tilgang til viktige private og offentlige tjenester
- Lovutøvelse
- Asyl, migrering og grensekontroll
- Administrasjon av rettigheter og demokratiske prosesser



Brussels, 21.4.2021
COM(2021) 206 final

ANNEXES 1 to 9

ANNEXES

to the

Proposal for a 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

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}

https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206

Risikostyring for høyrisiko AI systemer

Risikostyring inkluderer å:

- Identifisere **farer og uønskede hendelser** som en må kunne forvente ved bruk av AI systemet
- **Estimere og evaluere risikoen** – både for normal bruk og feilbruk
- Implementere **risiko-reduserende tiltak** i alle relevante faser livsløpet
- Avdekke **nye uønskede hendelser og farer** ved bruk av data samlet inn for AI systemet i bruk

Vi kjenner igjen Haddon's prinsipper…

Prioritert rekkefølge for risiko-reduserende tiltak

1. Tiltak som **fjerner** risikoen for uønskede hendelser
2. Tiltak som **begrenser** skaden forbundet med risikoen
3. Tiltak som **informerer** brukeren om risikoen og som ivaretar nødvendig opplæring og trening

Tillitsverdig (Trustworthy) AI

Viktig å enes om hva dette er

Initiativ fra DNV

1. Nøyaktighet/presisjon (accuracy)
2. Robusthet
3. Forklarbarhet (explainability)
4. Fortolkbar (interpretability)
5. (Fravær av) skjevhet (Bias)
6. (Ivaretagelse) av personvern (Privacy)
7. (Ivaretagelse) av IKT-sikkerhet (Security)

Kilde: DNV retningslinje **DNV RP 0771 (**under utgivelse)

Utgitt av EU

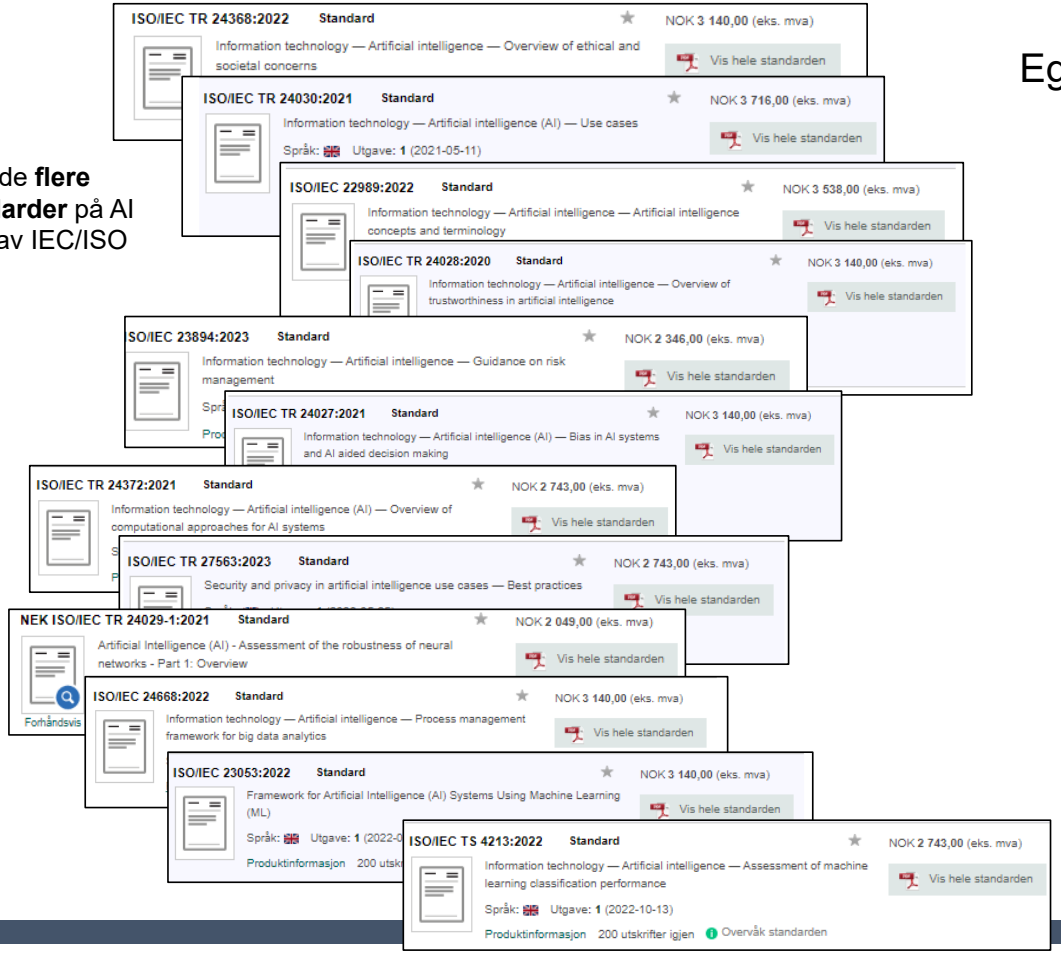
1. Rette seg etter lover og regler
2. Følge etiske prinsipper og verdier
3. Være robust mot utilsiktet skade

1. Tilpasset menneskets evner og behov for å forstå/påvirke/gripe inn (human agency and oversight)
2. Teknisk robusthet og sikkerhet, inkl. IKT-sikkerhet
3. Personvern og datahåndtering
4. Transparent (sporbart og forklarbart)
5. Rettferdig og ikke-diskriminerende
6. Bidra til samfunnet og miljøets beste
7. Å stole på og etterrettelig (accountability)

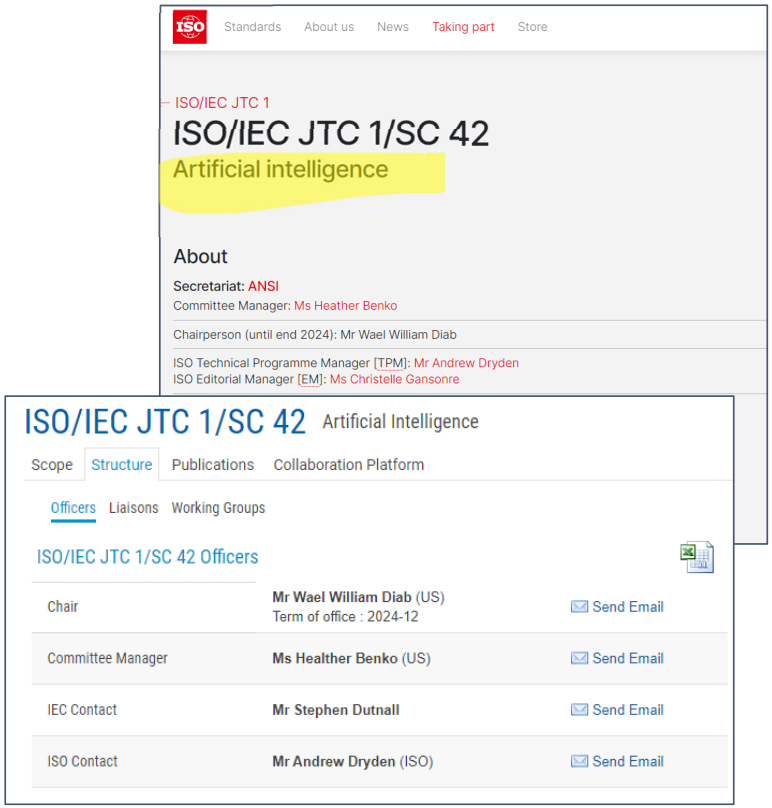
Kilde: **EU Ethics guidelines for trustworhty AI**

AI og internasjonal standardisering – masse foregår!

Allerede **flere standarder** på AI utgitt av IEC/ISO



Egen arbeidsgruppe felles for ISO og IEC



AI og funksjonell sikkerhet: IEC TR 5469 (utkast)

65A/1044/DC

For IEC use only

2022-06-24

INTERNATIONAL ELECTROTECHNICAL COMMISSION

TECHNICAL COMMITTEE 65: INDUSTRIAL-PROCESS MEASUREMENT, CONTROL AND AUTOMATION

SUB-COMMITTEE 65A: SYSTEM ASPECTS

Document for comment:

DTR document ISO/IEC TR 5469 'Artificial intelligence — Functional safety and AI systems' - for SC 65A comment

Following the circulation of 65A/1001/DC and the resulting 65A/1009/INF, ISO/IEC TR 5469 has now progressed to the final DTR ballot under ISO/IEC JTC 1/SC 42.

As part of our liaison with ISO/IEC JTC 1/SC 42, this is being circulated to SC 65A members **for comment only** (no vote).

Please see attached the DTR document of ISO/IEC TR 5469 for comment.

Comments received will be passed on to ISO/IEC JTC 1/SC 42/WG 3 for consideration, noting that following a DTR ballot only editorial comments can be considered.

National committees are requested to send their comments on the IEC commenting template to the IEC Central Office using the electronic voting system by no later than

2022-09-23

Ikke publisert ennå, men under arbeid

Hva er problemet (eller utfordringen)

Et SIS system er kjennetegnet av at:

- Funksjonalitet er tydelig spesifisert
- Logiske operasjoner er forklarbare og (med høy nok grad) forutsigbare
- Spesifisert funksjonalitet er testet og ofte også sertifisert
- Stabilitet i funksjonalitet som muliggjør å beregne pålitelighet (SIL) basert på data
- AI er ikke lov…

I dag fraråder IEC 61508 bruk av AI … I utkast til ny versjon av IEC 61508 står det at de ikke vil fraråde bruk av AI ifbm sikkerhetssystemer , men at utvikling av software for AI er utenfor scope og implementeres i hht IEC/ISO TR 5469. Hardware som implementerer AI vil måtte følge krav i IEC 61508 (og få SIL krav)

Et AI system kjennetegnes av at:

- Ikke all funksjonalitet i programvaren er beskrevet eller synlig
- Programvaren kan feile på nye og uventede måter
- Ny funksjonalitet kan utvikles over tid
- …

AI og sikkerhetskritiske systemer

Locks for the use of IEC 61508 to ML safety-critical applications and possible solutions

Albin Tarrisse

QRIB, French National Institute for Industrial Environment and risks (Ineris), France. E-mail: albin.tarrisse@ineris.fr

François Massé

QRIB, French National Institute for Industrial Environment and risks (Ineris), France. E-mail: francois.masse@ineris.fr

In the past few years, Artificial Intelligence (AI) and more precisely Machine Learning (ML), have been developing quickly and have found applications in a lot of various fields, including industry. The process industry and the use of machinery are critical because they represent a high risk for the surroundings, humans, goods and environment. To mitigate those risks, safety devices carrying out safety functions are used to monitor and control processes and machinery in real time. IEC 61508 which gives a framework to certify the Safety Integrity Level (SIL) of such systems is not suitable for complex algorithms like ML algorithms can be. We will attempt in this paper to discuss the general concepts of the standard that hinder the development and assessment of a safety devices using AI in the current state of knowledge. To the end, after describing the specificities of safety-critical systems and presenting the basic principles of AI and ML, this paper highlights the locks for the use of IEC 61508 and suggests some leverages, such as the use of a new safety lifecycle, a minimal level of transparency and explainability as well as the demonstration of robustness based on formal verification, testing and new metrics.
Disclaimer: AI is an area of research which is currently booming, the scientific knowledge is progressing rapidly, especially on the topics of safety assurance, testing, V&V, robustness, explainable AI (XAI). Thus, this article could not be exhaustive and represents our current state of knowledge on that matter. Some information in this paper have been taken from standards projects of the ISO/IEC JTC 1 SC42.

Keywords: Functional safety, Software Reliability, Process Industry, Machinery, Machine Learning, Artificial Intelligence.

1. Introduction

Artificial intelligence (AI) in its broadest sense have been developing in a theoretical and in an applicative point of view for more than 70 years now. The recent technical progresses, such as: the improvement of the collect and storage of large amount of data, the increase of computing capacity, and the development of new mathematical models have sped up the expansion of AI and more specifically of machine learning (ML) to such extent that it deeply affects our society. To that matter, the industry through the ongoing development of Industry 4.0, is no exception. The process industry and the manufacturing industry using machinery represent a high risk for the surroundings, humans, goods and environment. To mitigate those risks, safety devices carrying out safety functions are used to monitor and control processes and machinery in real time. The standard IEC 61508 gives a framework to qualify the risk reduction of Electrical/Electronic/Programmable Electronic Safety-related devices by assessing the functional safety. However, the standard as it stands is not suitable for complex algorithms such as ML. After introducing background knowledge and specifying what is at stake about ML and functional safety, we provide an overview of the locks for the use of IEC 61508 to ML safety critical applications. Finally, from these general concepts, we suggest some opportunities and leverages that can be worked on to improve the assessment of the reliability of the safety-related devices.

2. Description of the devices

The devices covered by the study are constituted with electronics components and have an embedded software that must achieve at least one safety function. The safety function results from a need to reduce the risk due to a hazardous processes or machinery. It can either be a protection system or a control system.

They can be part of a greater safety system which is usually made of: one or several sensors that measure and process a physical signal, a logic which encompasses all the operating rules, and one or several actuators that act given the measured signal and the logic in order to put and maintain the system under control in a safe state of operation.

As an illustration, it could be a sensor of various type that measures temperature, pressure, level, chemical concentration, voltage, power, radiation or that captures a sound, a picture or a video.

The safety devices considered in this paper are automated up to a certain extent, because they are operating under specific conditions without human intervention. There are

Table 1. Blocking points for using AI in safety critical systems

Phases	Activities	Properties
Specification	Formal Method	Completeness and correctness
Specification	Forward traceability between the system safety requirements and the software safety requirements	Completeness
Design and development	Certificated tools	Clarity, correctness, repeatability
Design and development	Design and coding standards; Structured programming	Fault avoidance, simplicity, predictability, testability
Design and development	Module testing	Completeness
Design and development	Test coverage	Completeness, Correctness
Design and development	Modular approach	Fault avoidance, simplicity, predictability, testability
Software Verification	Formal verification and formal proof	Correctness
Software Verification	Static and dynamic analysis	Completeness, Correctness

Kommer også med noen forslag til hvordan ML kan utvikles «sikkert»

Kilde: https://rpsonline.com.sg/proceedings/9789811820168/html/661.xml

IEC 61508 – Håndtering av AI (forslag)

IEC 61508-4: Nye definisjoner

- Software technology class I:
  Being developed by **human programming or coding**, **which can be** completely followed up, understood and reviewed in all lifecycle phases, including the software off-line support tools used (e.g. code compilers), where **all data sources / parameters** are **predetermined and limited** by humans and **no functional self-evolvement** of the software itself during operation is possible.

- Software technology class II:
  At least initially being developed by human programming but **partly developed or coded by software algorithm**s (e.g., machine learning), ***which cannot*** be completely... (same text as above)

- Software technology class III:
  Software not meeting criteria of I and II

Traditional

Machine learning

Natural language processing,…

IEC 61508 – Håndtering av AI (forslag)

Tydeliggjøring av at utvikling av AI algoritmer og modeller ikke dekkes av IEC 6108:

- IEC 61508-1, kap 1.2: Nytt punkt p) og IEC 61508-3, kap 1.2 Nytt punkt (k):

  *"**does apply** to software algorithms of software technology **class I** (see definition in Part 4 Clause TBS),but does **not apply** for software algorithms of software technology **class II and III** (see definitions in Part 4, Clause TBS and Clause TBS).*

  *NOTE 1 The software algorithms classes relate to the generically used term "artificial intelligence" and correspond to the activities of ISO/IEC JTG1, SC42, **IEC TR 5469:2023.***

  *NOTE 2 The concept of safety integrity level as described in this document does not fully apply to software of technology class II and III. Refer to **ISO/IEC TR 5469:2023** for further details.*

  *NOTE 3 The concept of **safety integrity level as described in this document applies to the hardware** used to execute or implement software algorithms of software technology class I, II or III*

Klassifisering av AI i SIS (forslag i IEC TR)

Bruksnivåer (usage levels)

MÅ se en gang til på denne!

SIS does automated decision-making	**AI part of SIS**			
	Yes	No	No, but AI part of develoment tool	No, but potentially impacted by another AI system
Yes, with AI	A1	A2		
Yes, no AI			B1	C
Not automated		D	B2	C

Mulighet for å bruke eksisterende standarder for funksjonell sikkerhet (eks: IEC 61508, ISO 26262,..)

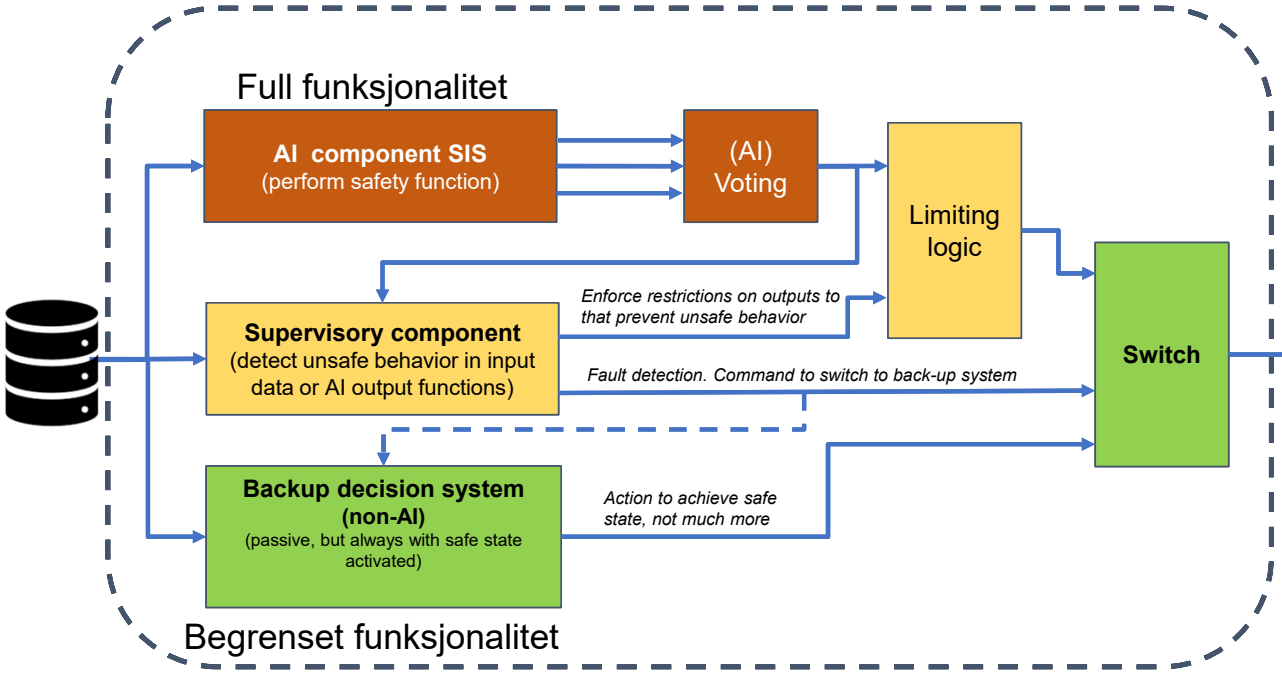
- Class I: AI teknologien kan utvikles innenfor eksisterende funksjonell sikkerhetsstandard (☺)
- Class II: AI teknologien kan delvis utvikles innenfor eksisterende funksjonell sikkerhetsstandard (😐)
- Class III: AI teknologien kan ikke utvikles innenfor eksisterende funksjonell sikkerhetsstandard (☹)

Hvis noen tilleggskrav oppfylles

Ikke prøv (ennå)

Implementering av AI komponent i IEC/ISO TR 5469

SIS implementeres i flere lag – som må være uavhengig



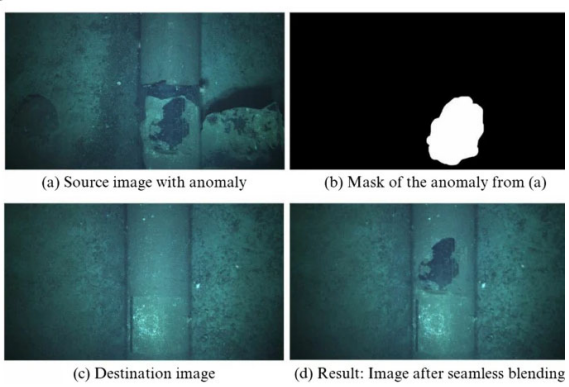
Output (sikker aksjon) – enten:

- Utført av AI komponent
- Modifisert av supervisory komponent
- Utført av backup system

AI i sikkerhetskritiske systemer: Spesielle utfordringer

Å ha nok relevante data for trening!

Mange anlegg er designet med SIS systemer som skal **operere sjelden.** Dermed lite eller ingen data som bekrefter oppførsel og situasjoner som ledet til at SIS systemet måtte «trå til». Kritiske tilstander man nå oppdage kan hende skjer veldig sjelden.



Simulatorer kan brukes til å generere scenariene for å teste responsen fra AI-SIFer, men det vil alltid hefte usikkerhet om disse representerer «virkeligheten».

Genering av syntetiske data kan brukes til å generere ekstra treningsdata ifbm klassifisering av kritiske tilstander, men de må være realistiske og merkes.

https://link.springer.com/article/10.1007/s44163-023-00069-1

AI i sikkerhetskritiske systemer: Andre utfordringer

Diskuterer teknikker og metoder som kan forhindre:

- Utilsiktede negative konsekvenser i oppfyllelse av målet som er satt for det autonome systemet (robotstøvsuger som ødelegger ledninger på gulvet)
- Utilsiktet effekt av valgte belønningskriterier («reward hacking») (robotstøvsuger slår bare av kamera så den ikke ser ledningene på gulvet. Ute av sinn…)
- Uønskede handlinger som er resultat av mangel på trening i sjeldne situasjoner som er kostbare å trene på (scalable oversight) (robotstøvsuger bør klare å stoppe opp når den kommer over garnnøster, selv om det skjer sjelden)
- Utrygg utforskning av miljøet for å lære (robotstøvsuger kjører utfor trapp)
- Tap av robusthet mot endring i distribusjon/fordelinger i data (robotstøvsuger oppfatter ikke at den er blitt plassert til å jobbe i en hønsegård)

Fokus på reinforced (forsterket) læring og autonome systemer

Concrete Problems in AI Safety

Dario Amodei* Google Brain
Chris Olah* Google Brain
Jacob Steinhardt Stanford University
Paul Christiano UC Berkeley
John Schulman OpenAI
Dan Mané Google Brain

Abstract

Rapid progress in machine learning and artificial intelligence (AI) has brought increasing attention to the potential impacts of AI technologies on society. In this paper we discuss one such potential impact: the problem of *accidents* in machine learning systems, defined as unintended and harmful behavior that may emerge from poor design of real-world AI systems. We present a list of five practical research problems related to accident risk, categorized according to whether the problem originates from having the wrong objective function ("avoiding side effects" and "avoiding reward hacking"), an objective function that is too expensive to evaluate frequently ("scalable supervision"), or undesirable behavior during the learning process ("safe exploration" and "distributional shift"). We review previous work in these areas as well as suggesting research directions with a focus on relevance to cutting-edge AI systems. Finally, we consider the high-level question of how to think most productively about the safety of forward-looking applications of AI.

1 Introduction

The last few years have seen rapid progress on long-standing, difficult problems in machine learning and artificial intelligence (AI), in areas as diverse as computer vision [82], video game playing [102], autonomous vehicles [86], and Go [140]. These advances have brought excitement about the positive potential for AI to transform medicine [126], science [59], and transportation [86], along with concerns about the privacy [76], security [115], fairness [3], economic [32], and military [16] implications of autonomous systems, as well as concerns about the longer-term implications of powerful AI [27, 167].

The authors believe that AI technologies are likely to be overwhelmingly beneficial for humanity, but we also believe that it is worth giving serious thought to potential challenges and risks. We strongly support work on privacy, security, fairness, economics, and policy, but in this document we discuss another class of problem which we believe is also relevant to the societal impacts of AI: the problem of accidents in machine learning systems. We define accidents as unintended and harmful behavior that may emerge from machine learning systems when we specify the wrong objective function, are

*These authors contributed equally.

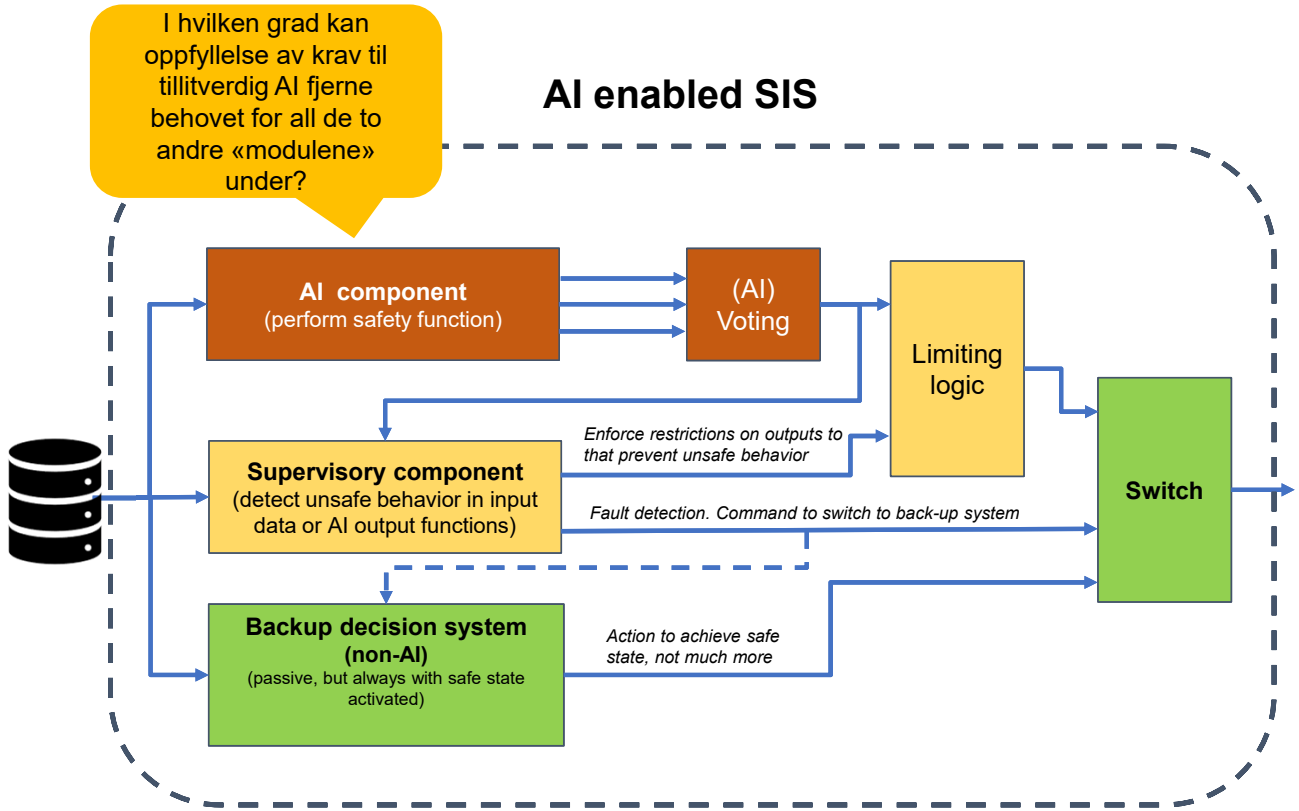
https://arxiv.org/abs/1606.06565

Tilleggskrav for SIS med AI (eksempler)

Oppfylle krav til tillitsverdig AI (EU, DNV) – se egen liste

I tillegg:

- Kompensere for uønsket funksjonalitet vha arkitektur (IEC TR 5469)
- Undersøke også fysiske årsaker som bortfall av viktige sensorer, utskiftning av utstyr,..
- Gjenta treningen med jevne mellomrom



Klarte vi å belyse motivasjonen?

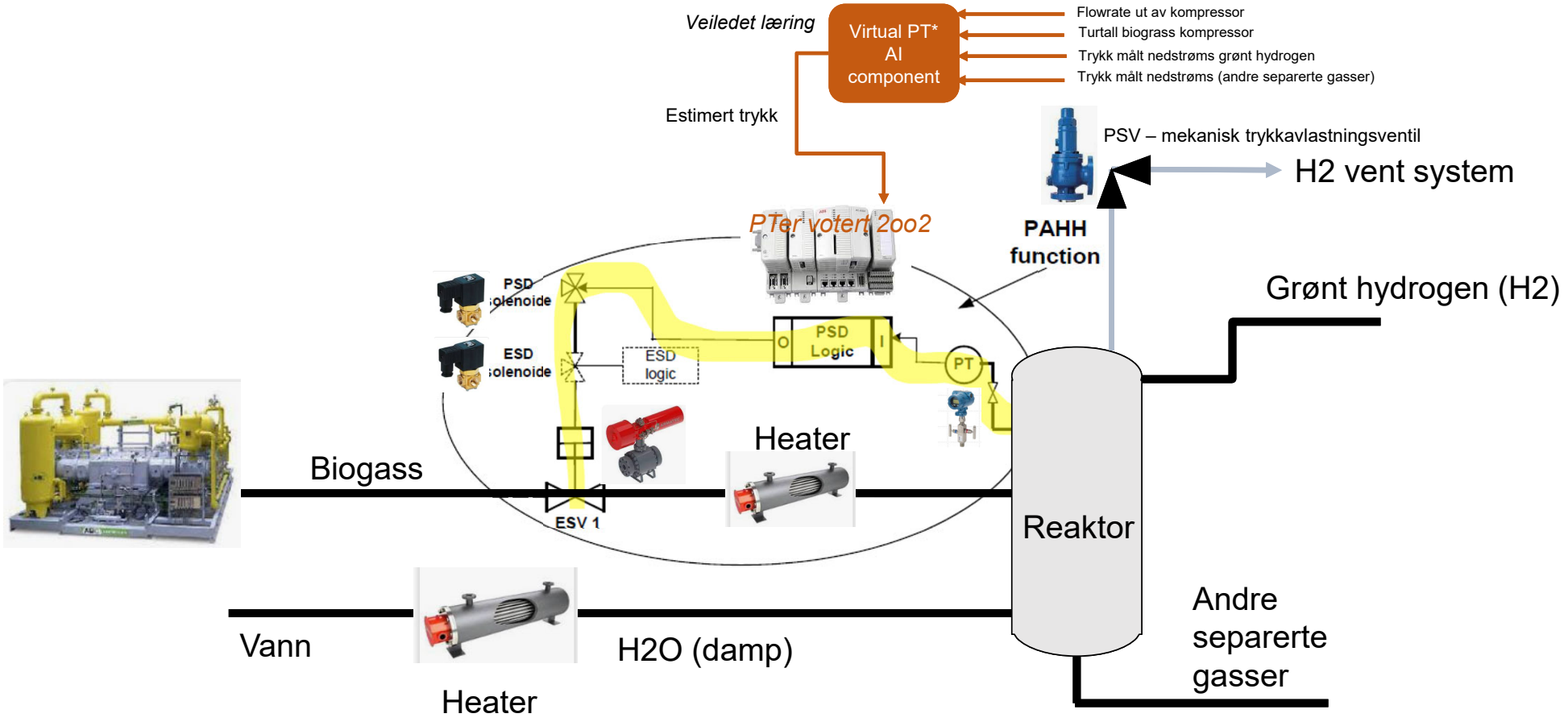
Motivasjon

I dag **tillates** normalt **IKKE** kunstig intelligens industrielle sikkerhetskritiske systemer. Samtidig bør kunstig intelligens (KI/AI) ha **potensialet** til å bidra til **økt sikkerhet**. Men **uforutsigbarheten** i oppførselen og **usikkerheten om evnene** til KI/AI systemer må reduseres og funksjonsdyktigheten må kunne dokumenteres.

Vi må derfor finne måter å bruke kunstig intelligens på en trygg måte også for sikkerhetskritiske systemer.

Gruppeoppgave 3

Case SIF (gjennomgående eksempel)



Gruppeoppgave 3 (mellom seminar 3 og 4)

1. Beskriv kort en utvidelse av SIFen dere har valgt med en «AI-komponent»
2. Ta for dere alle boksene (modulene) i figuren til høyre og gi eksempler på hva de kan inneholde gitt deres case
3. Beskriv kort kravene til «Tillitsverdig AI» for høyrisikosystemer og kommenter kort på hvorfor en arkitektur som den til høyre allikevel kan være nødvendige selv om kravene er oppfylt
4. Identifiser eksempler på ulike typer bias-kategorier som ble presentert i seminaret.