DNV

WHEN TRUST MATTERS

Assurance of AI-enabled systems

AI+, Halden - Norway

Christian Agrell

03 May 2023

A global assurance and risk management company

159
years

12,000
employees

100,000
customers

100+
countries

5% R&D
of annual revenue

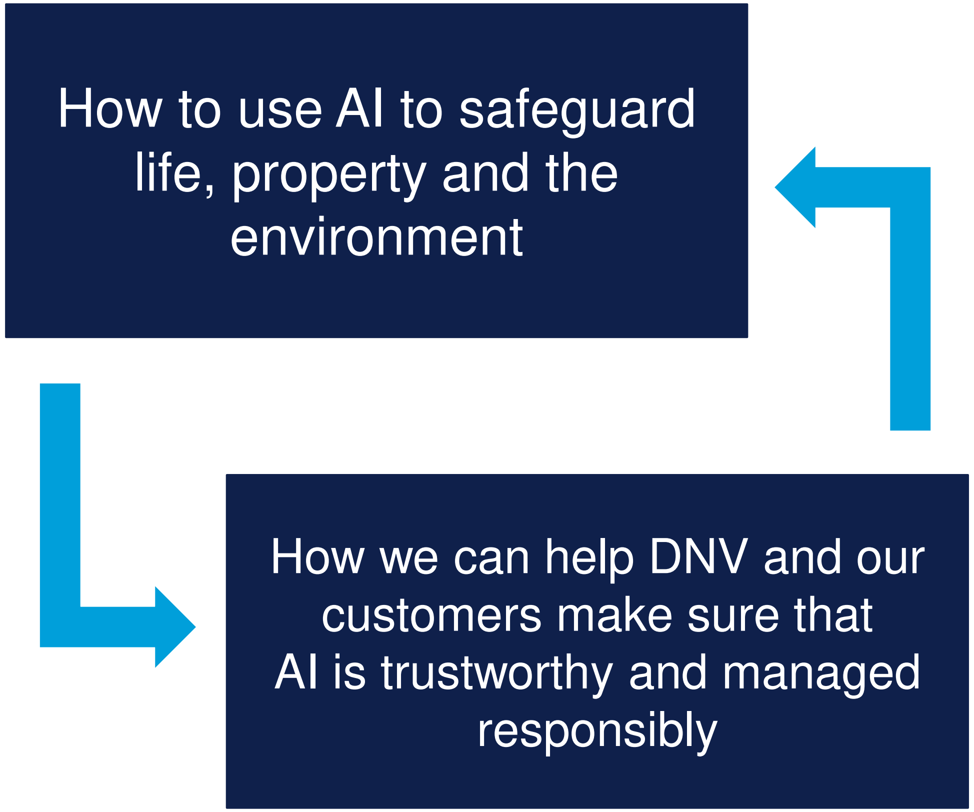
Ship and offshore classification and advisory

Energy advisory, certification, verification, inspection and monitoring

Management system certification, supply chain and product assurance

Software, platforms and digital solutions

AI research at DNV



How to use AI to safeguard life, property and the environment

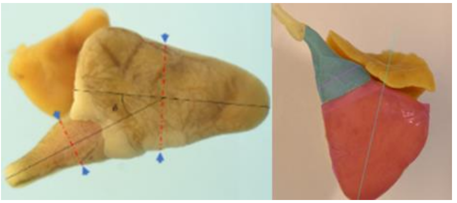
How we can help DNV and our customers make sure that AI is trustworthy and managed responsibly

Inspection

- Using computer vision to detect cracks, corrosion, deformations, etc.

Predictive maintenance and health monitoring

- Incl. solar, wind, oil & gas, maritime and aquaculture



Autonomy

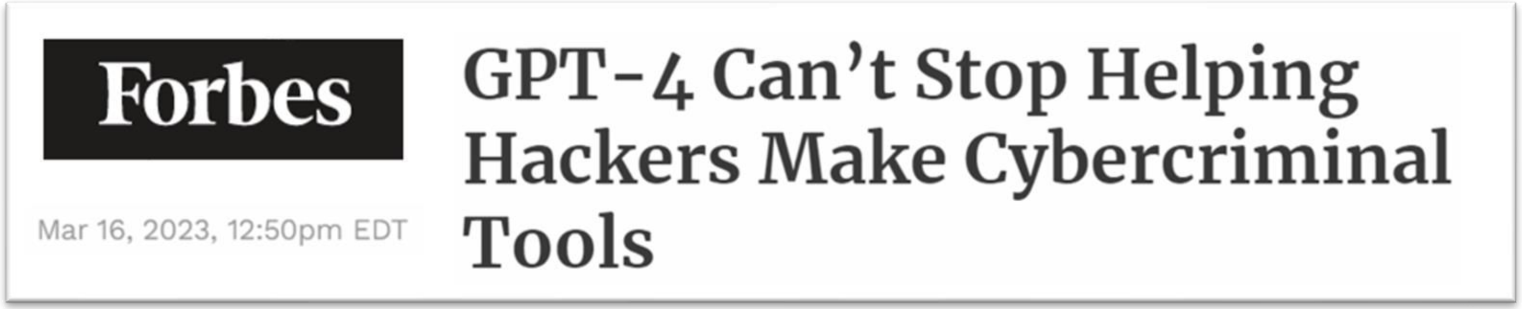
- Situational awareness
- Safe reinforcement learning

Assurance of AI-enabled systems

- Demonstrate that a certain application of AI will be sufficiently safe, reliable, fair, transparent, etc.

WHY
do we need assurance of AI?

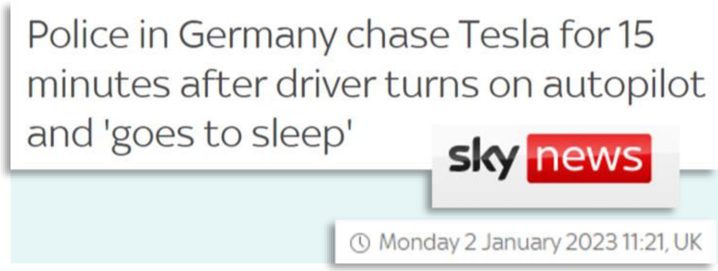
What are the consequences of using AI ?



Forbes

Mar 16, 2023, 12:50pm EDT

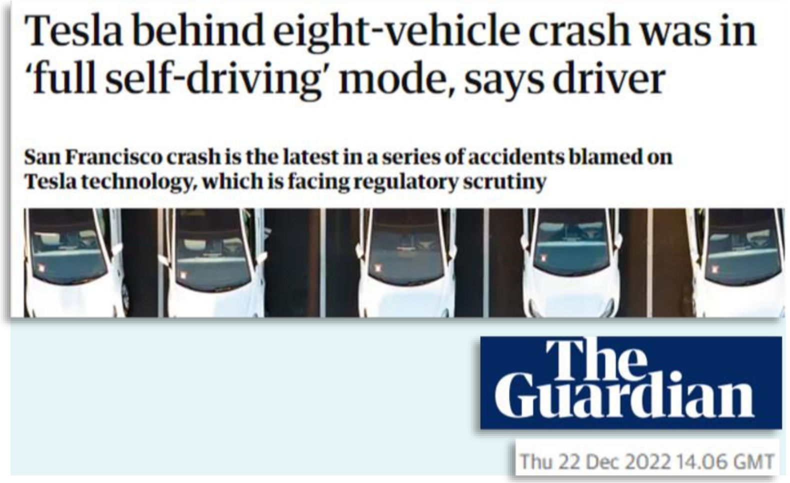
GPT-4 Can't Stop Helping Hackers Make Cybercriminal Tools



Police in Germany chase Tesla for 15 minutes after driver turns on autopilot and 'goes to sleep'

sky news

Monday 2 January 2023 11:21, UK

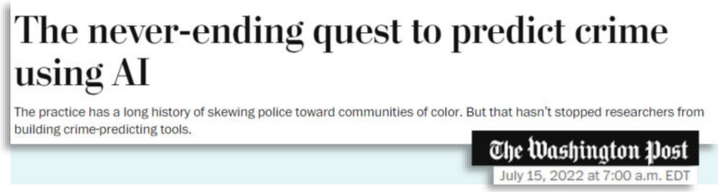


Tesla behind eight-vehicle crash was in 'full self-driving' mode, says driver

San Francisco crash is the latest in a series of accidents blamed on Tesla technology, which is facing regulatory scrutiny

The Guardian

Thu 22 Dec 2022 14.06 GMT

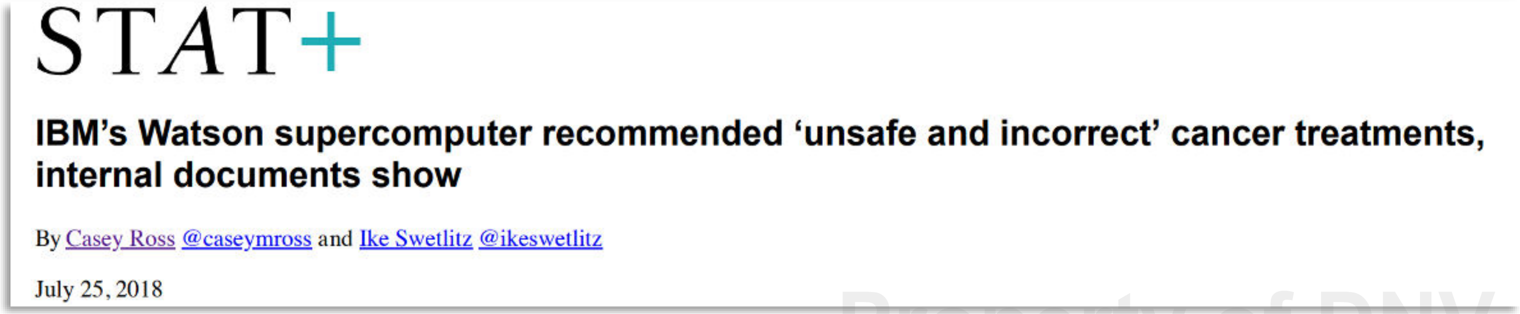


The never-ending quest to predict crime using AI

The practice has a long history of skewing police toward communities of color. But that hasn't stopped researchers from building crime-predicting tools.

The Washington Post

July 15, 2022 at 7:00 a.m. EDT



STAT+

IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show

By Casey Ross @caseymross and Ike Swetlitz @ikeswetlitz

July 25, 2018

DNV

AI has a vast potential to advance business, improve lives and tackle global challenges.

Business Intelligence

Autonomy

Science

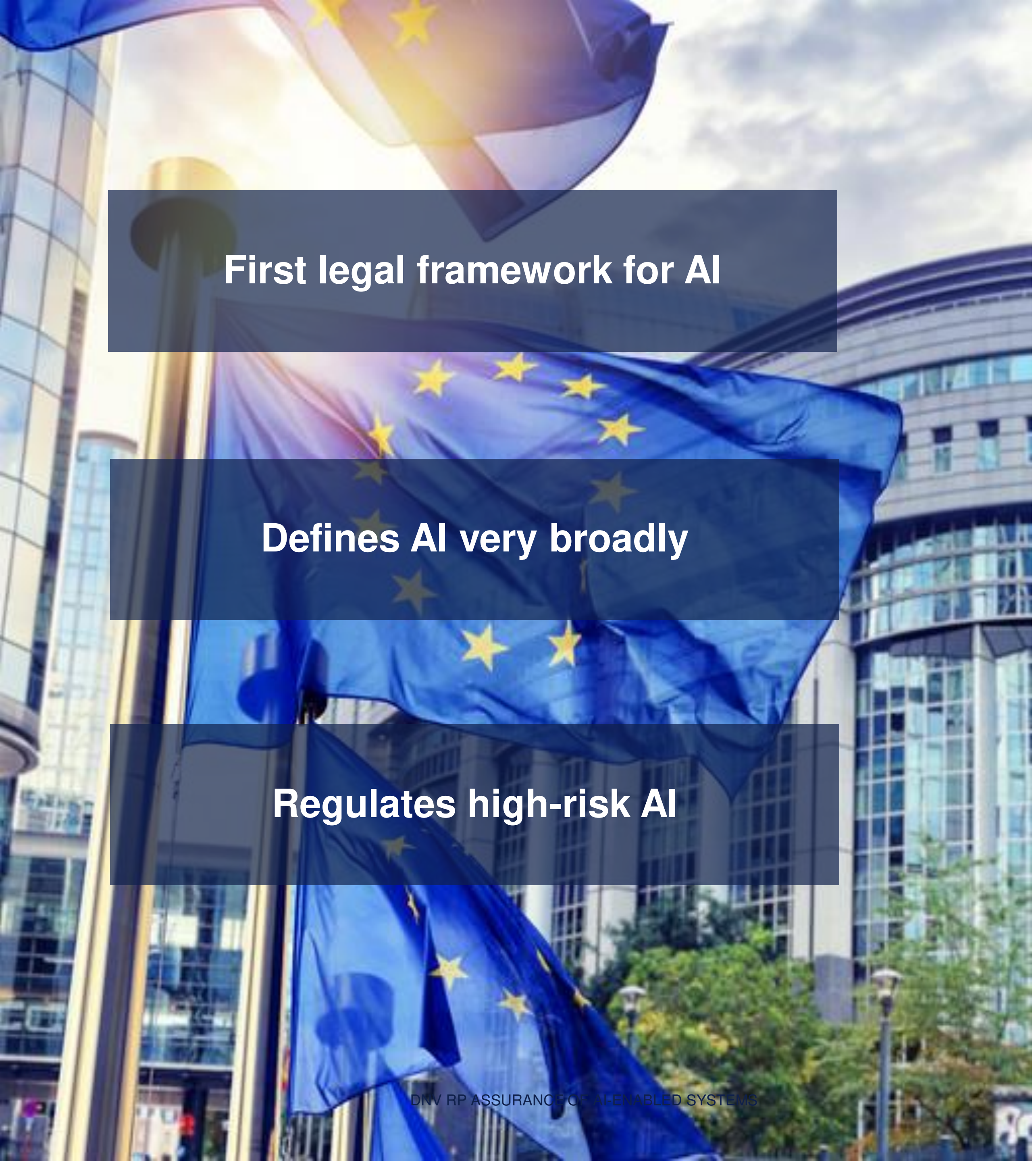
But trust is needed for this to materialise

- 61% of users are either ambivalent or unwilling to trust AI.

  [KPMG and University of Queensland 2023, Trust in Artificial Intelligence - A global study. (17 countries, 17.000 participants)]

- 84% of IT professionals now saying that being able to explain how their AI arrives at different decisions is important to their business

  [IBM Global AI Adoption Index 2022 (13 countries, 7 502 participants)]



First legal framework for AI

Defines AI very broadly

Regulates high-risk AI

The EU AI Act

- The AI Act will pass EU Council in 2023
- 2 year “grace period” (as with GDPR)

*Goal: Foster the development, use and uptake of AI in Europe, by ensuring **trustworthy** and **responsible** AI*

WHAT

is assurance of AI?

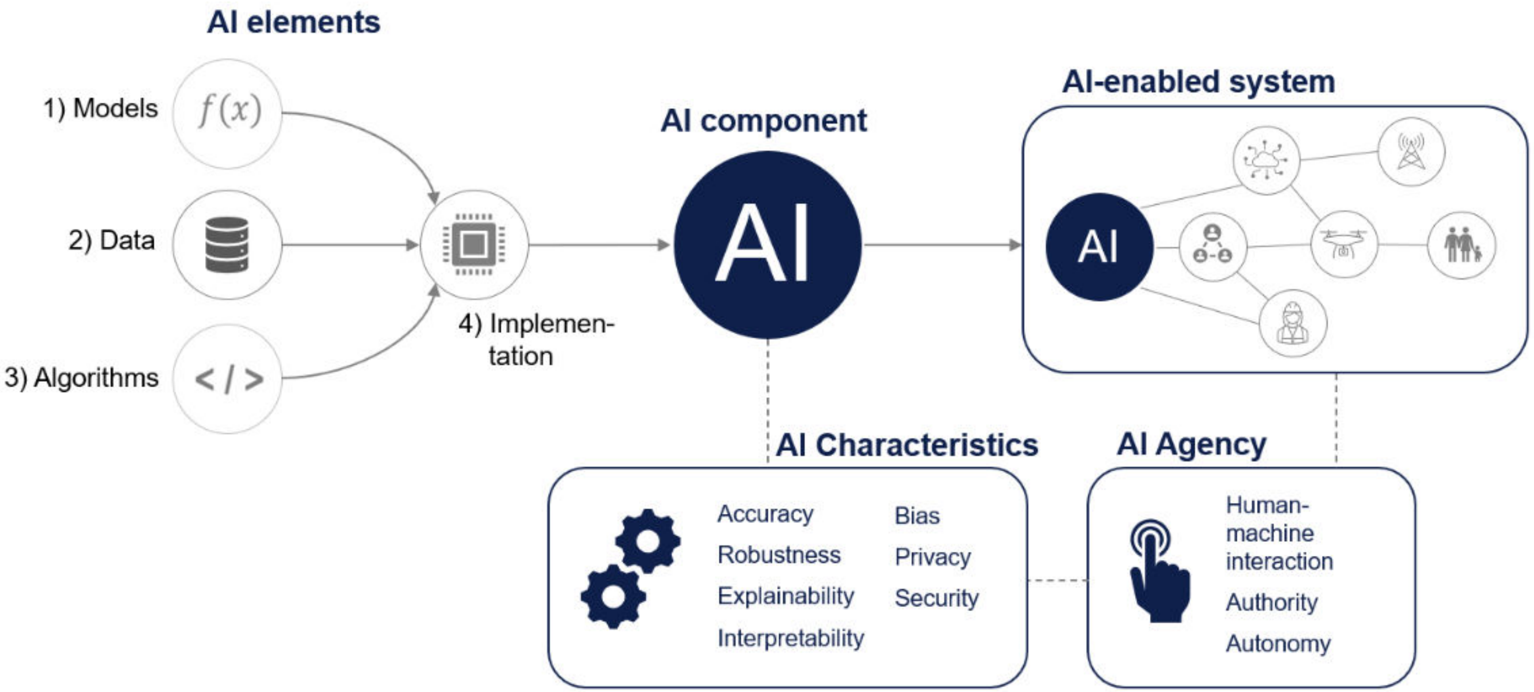
Assurance:

Grounds for justified confidence that a claim has been or will be achieved (ISO 15026-1)

Example of claim:

The **ship** is sufficiently **safe**

AI-enabled system



AI elements

1) Models

2) Data

3) Algorithms

4) Implementation

AI component

AI

AI-enabled system

AI

AI Characteristics

Accuracy

Robustness

Explainability

Interpretability

Bias

Privacy

Security

AI Agency

Human-machine interaction

Authority

Autonomy

Claims

Some elements of **trustworthy AI** formulated as assurance **claims**:

- The **system** is sufficiently ***safe***
- The **system** is sufficiently ***robust***
- The **system** is sufficiently ***accurate***
- The **system** is sufficiently ***interpretable***
- The **system** is sufficiently ***transparent***
- The **system** is sufficiently ***explainable***
- The **system** is sufficiently ***fair***
- The **system** is sufficiently ***secure***

HOW

can we perform assurance of AI?

1) Dealing with **complexity** and **emergence**

- Examples of **complex systems**: Traffic flows, financial markets, the earth's climate, pathogens, ecosystems, the internet.

- Examples of **emergent** behaviour:

Building new highways to reduce traffic congestion → attracts new drivers → more congestion

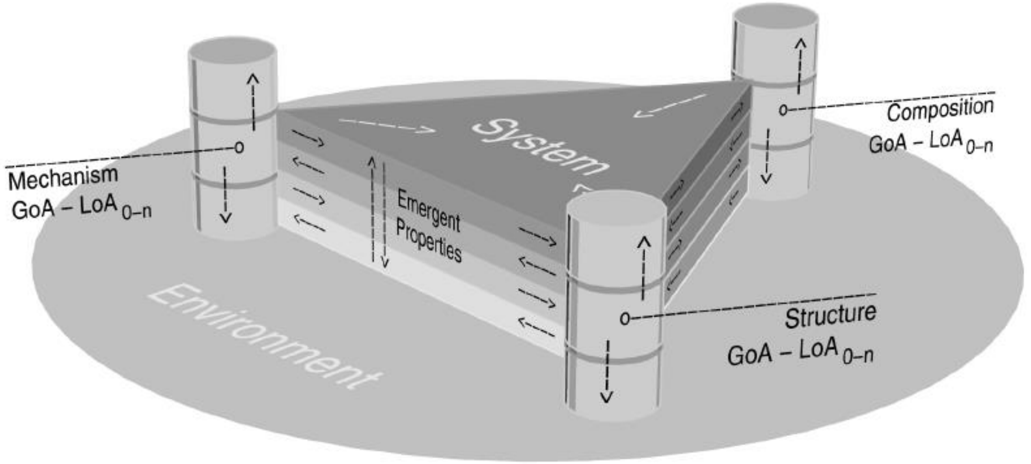
Autonomous cars to increase safety → pedestrians have no fear of cars and "rule the streets" → cars cannot move

1) Dealing with **complexity** and **emergence**

Emergent properties
are properties that become apparent and result from various interacting components within a system but are properties that do not belong to the individual components themselves.

We can deal with complexity and emergence through a **systems approach**

- AI-enabled systems are often **complex**
- Safety, fairness, transparency, explainability and interpretability are ***emergent* properties**



[Figure from the book **Demonstrating safety of software-dependant systems**, Editors: Meine van der Meulen and Tore Myhrvold, 2022]

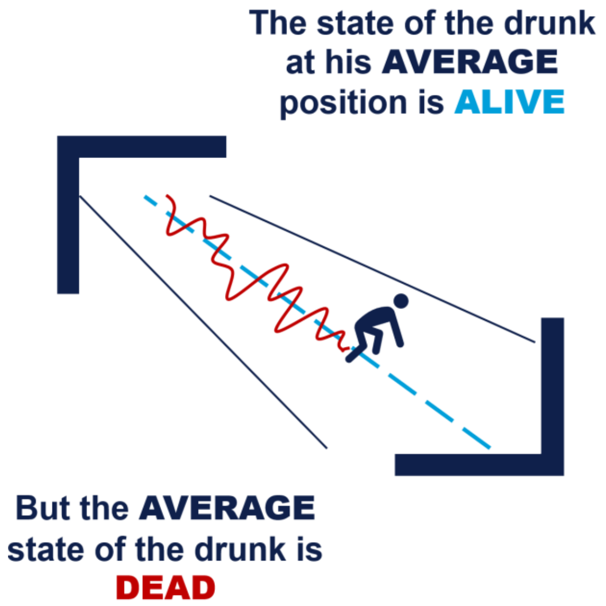
2) Dealing with **uncertainty**

Humans are bad at reasoning about probability and uncertainty.

Anna is a very structured, a little shy and has a passion for reading books. Is it most likely that Anna a librarian or a farmer?

Shuffle a deck of cards. What is the chance that there has existed a deck of card in the same order, ever?

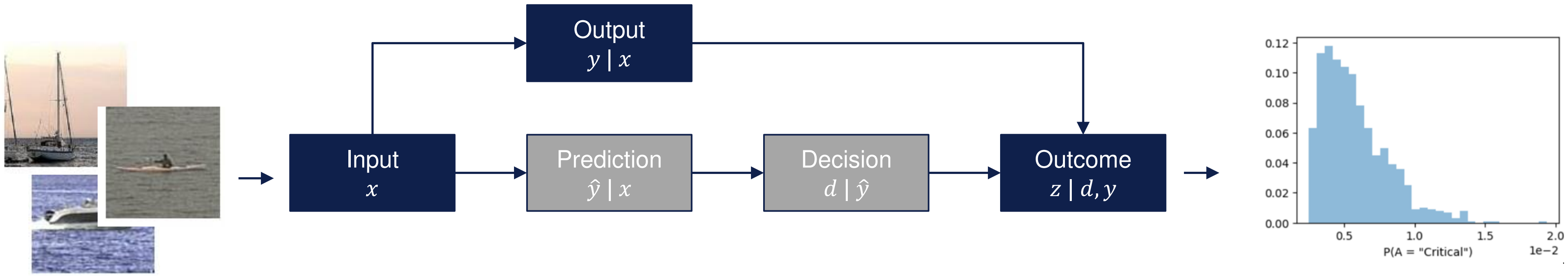
The effect of uncertainty (risk) is important


The state of the drunk at his **AVERAGE** position is **ALIVE**
But the **AVERAGE** state of the drunk is **DEAD**

Decision Making with Insight 2nd Edition, Sam L. Savage, 2003

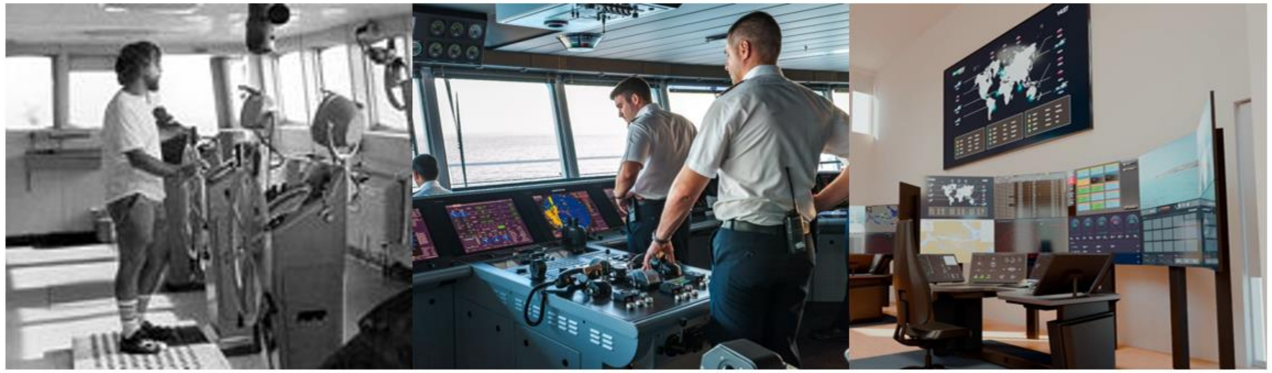
2) Dealing with **uncertainty**

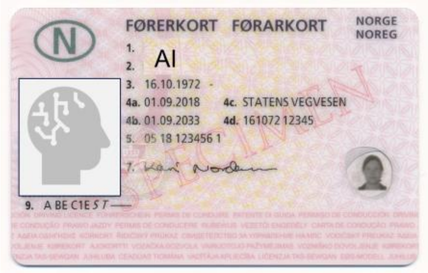
- Machine learning models introduce uncertainty
- We need to understand the effect of this uncertainty (risk)
- This requires propagation of uncertainty between components and sub-systems



3) Dealing with **novelty**

New technology comes with new unknowns





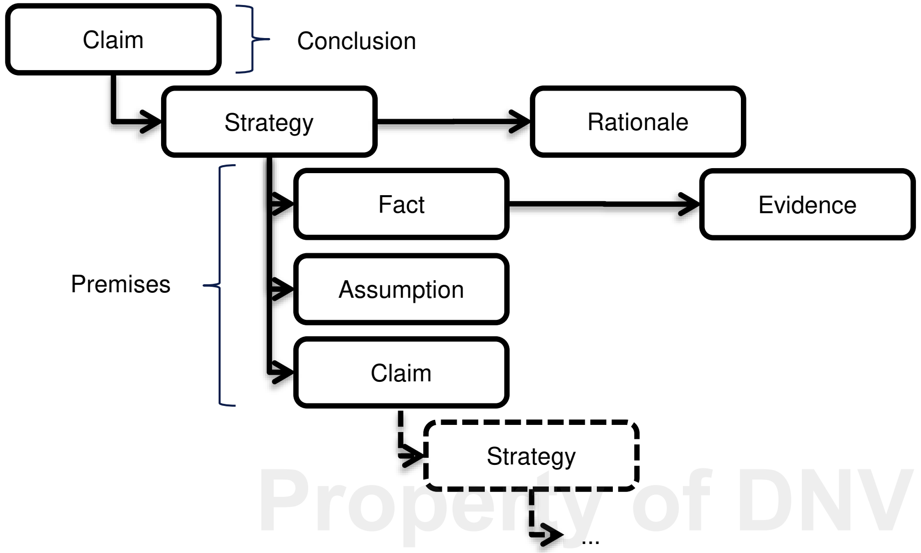
FØRERKORT FØRARKORT

NORGE
NOREG

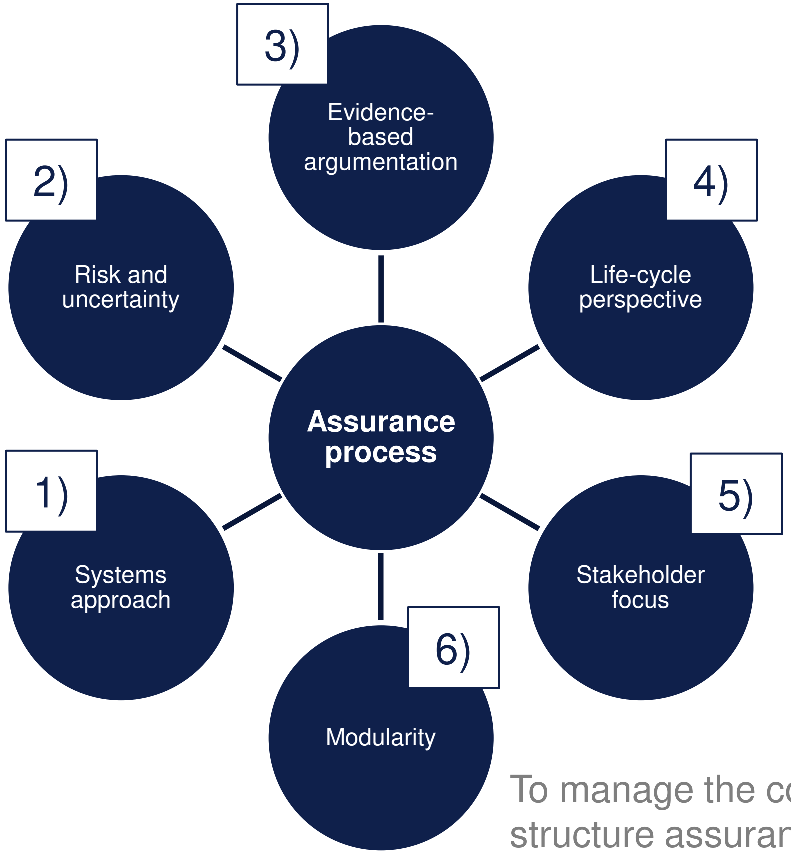
1.
2. AI
3. 16.10.1972 -
4a. 01.09.2018 4c. STATENS VEGVESEN
4b. 01.09.2033 4d. 161072 12345
5. 05 18 123456 1
7.
9. A BE C1E S T —

Why can't we use our current driving tests to give an autonomous car a licence?

Evidence-based reasoning is a structured way of conducting assurance, which is applicable for **qualification of new technology**



Key concepts of our assurance process



Assurance process

1) Systems approach

To create an understanding of how AI impacts emergent behaviour on system level, to guide how assurance on system level requires assurance of the AI

2) Risk and uncertainty

To guide what needs to be assured based on the presence of uncertainty and knowledge gaps and potential effects on stakeholders' interests

3) Evidence-based argumentation

To structure and present relevant knowledge about the system in scope

4) Life-cycle perspective

To ensure that stakeholder interests are respected throughout the entire life cycle of the AI enabled system

5) Stakeholder focus

To place focus on what matters to stakeholders

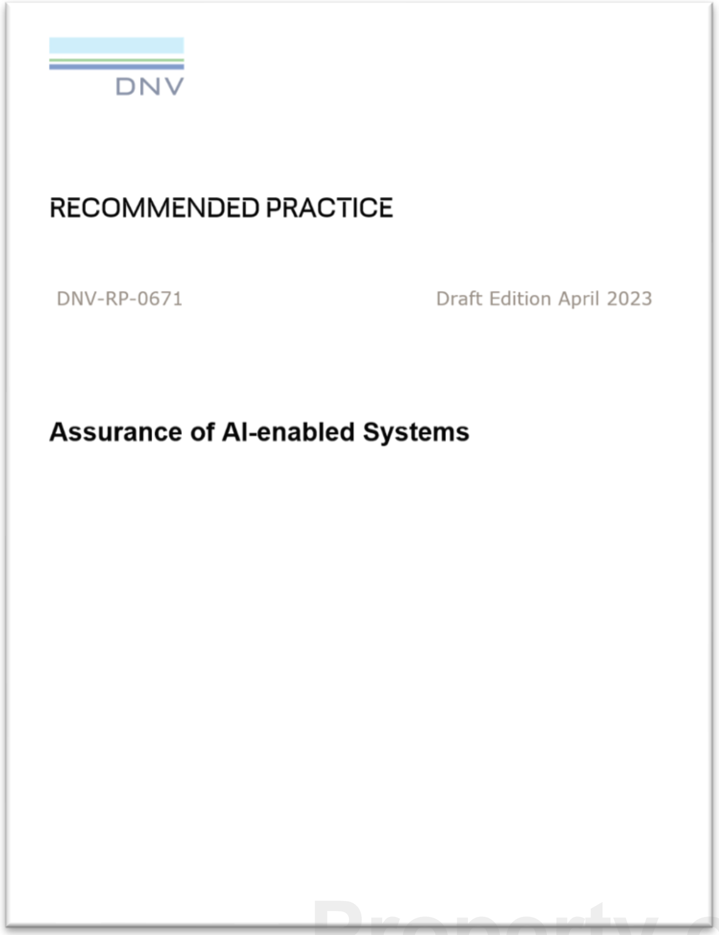
6) Modularity

To manage the complexity of AI-enabled systems, to structure assurance arguments, and to facilitate collaboration among stakeholders

New DNV Recommended Practice (RP)

- A new Recommended Practice **DNV-RP-0671 Assurance of AI-enabled systems** is on its way
- Builds the 6 key concepts into a process
- Gives guidance on how to ensure that AI is **trustworthy** and managed **responsibly**
- Can be used for compliance with the **EU AI Act**

- The RP goes on external hearing May 2023, and will be available to the public later in 2023



DNV

RECOMMENDED PRACTICE

DNV-RP-0671 Draft Edition April 2023

Assurance of AI-enabled Systems

WHEN TRUST MATTERS

Thank you for your kind attention!

Visit www.dnv.com/research

www.dnv.com

DNV