

Tillit til kunstig intelligens: Perspektiver og praksis

Forskningsparken, Oslo Science Park
13. november 2025

Velkommen!

Seminaret er i regi av *Gemini Centre for Trustworthy and Sustainable AI*

Arrangert av Sagar Sen og Gencer Erdogan, SINTEF Digital

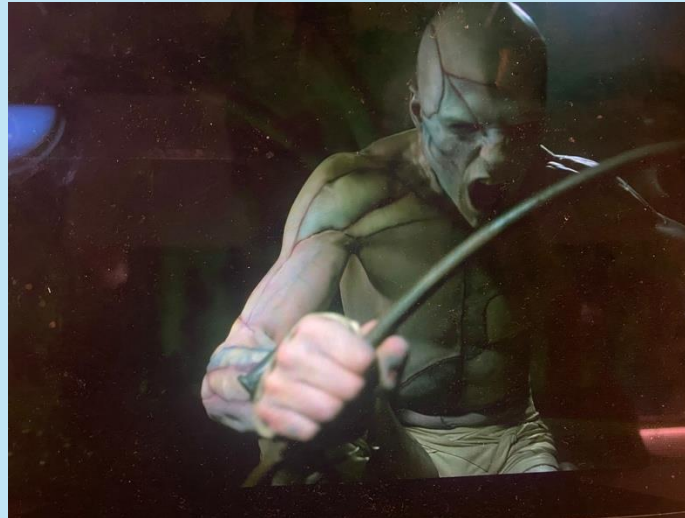
*“Trust is the invisible infrastructure that makes technology human.” – Onora
O’Neill (2002)*

Frankenstein gjenoppstår

Maktens drøm



Speilopprøret



Sammenbrudd og erkjennelse



Kilde: Frankenstein (2025), Netflix

Tillit før KI

Før KI formet tillit alle menneskelige relasjoner:

- **Med mennesker:** ærlighet og pålitelighet
- **Med dyr:** gjensidig læring og instinkt
- **Med planter:** tålmodighet, omsorg og gjensidighet – vi stoler på at det vi pleier, vil vokse
- **Med økosystemer:** gjensidig avhengighet bygget på sykluser, ikke kontroll
- **Med maskiner:** biler, verktøy og teknologier opparbeidet tillit gjennom konsistens
- **Med institusjoner:** tillit opprettholdt gjennom rettferdighet og ansvarlighet

“We learn to trust not perfection, but predictable imperfection.”
(Baier 1986; Luhmann 1979; Latour 1992)

Tillitens natur

“Trust is a bet on the future contingent actions of others.”

– Luhmann (1979)

Tillit bygger bro mellom usikkerhet og felles mål.

I KI: forutsigbarhet + åpenhet + delte verdier.

Vår reise begynner med å omfavne usikkerheten – ikke fjerne den.

En kort historie om tillit i KI

Epoke	Fokus	Tillitsutfordring
1950–70-tallet	Symbolsk KI	“Kan maskiner tenke?” – Turing (1950)
1980–2000	Ekspertsystemer	Overkonfidens, «black-box»-bias
2010–	Dyp læring	Dataskjevhet, manglende forklarbarhet
2020+	Grunnmodeller	Styring, sikkerhet, alignment

Hver generasjon av KI har satt samfunnets tillit på prøve.
(Turing 1950; Floridi & Cowls 2019; Russell 2021)

Dimensjoner av pålitelighet

Tre prøvelser på vår reise:

- **Teknisk robusthet** – Kan systemet være sikkert og forklarbart?
- **Etisk forsvarlighet** – Kan det handle rettferdig og ansvarlig?
- **Juridisk samsvar** – Kan det overholde lover og regler?
- **Heltenes utfordring:** å balansere innovasjon med integritet.
(EU AI Act 2024; ISO/IEC 42001:2023)

Det globale landskapet for tillit til KI

Over hele verden søker nasjoner etter pålitelig KI – både som etisk nødvendighet og som konkurransefortrinn.

- EU: [Ethics Guidelines for Trustworthy AI](#) (2019)
- USA: [NIST AI Risk Management Framework](#) (2023)
- Kina: [Generative AI Measures](#) (2023)
- OECD: [Human-centered AI Principles](#) (2021)

“Trust has become a geopolitical and moral currency.”

Den norske konteksten

Norges høye sosiale tillit støtter ansvarlig innovasjon.

Digitale offentlige tjenester (Altinn, BankID) trives gjennom åpenhet.

Men høy tillit skaper også høye forventninger.

Nasjonal strategi for kunstig intelligens

“Med stor tillit følger stort ansvar.”
(World Values Survey 2022; Slettebakk & Bjerke 2023)

Stemmer og utfordringer langs veien

“Tillit er umulig uten inkludering.” – Timnit Gebru

“KI reflekterer vår politiske økonomi.” – Kate Crawford

“KI trenger ikke bare tilpasning, men ydmykhet.” – Yoshua Bengio

Reisen krever lytting – til kritiske, mangfoldige og globale perspektiver.

Qualia: Erfaringens indre dimensjon

1. *Qualia* (entall *quale*)– de subjektive kvalitetene ved erfaring, hvordan det kjennes å se rødt, høre en stemme, eller føle tillit
2. Begrepet introdusert av C. I. Lewis (1929), videre utviklet av Thomas Nagel (1974) “What it means to be bat?” og Frank Jackson (1982)
3. **Tanke eksperiment:** “Mary, en vitenskapskvinne som kjenner alle fysiske fakta om farger, lever i et svart-hvitt rom til hun ser rødt og oppdager hvordan det oppleves.” Mary the Color Scientist, Frank Jackson (1982)
4. Et *quale* kan ikke reduseres til proposisjoner, data eller beskrivelser
5. KI kan simulere mønstre, men ikke *føle* mening

Fellesskapet i salen

- **Forskning og akademia:** SINTEF, SINTEF Digital, OsloMet, UiO, NTNU, UiA, Inven2
- **Offentlig sektor og styring:** Datatilsynet, Digitaliseringsdirektoratet, Forbrukerrådet, DNV, Helsedirektoratet, NVE, Mattilsynet, Stortinget
- **Industri og teknologi:** Equinor, Telenor, Cognite, Kongsberg Gruppen, Atea, Accelerate at Iver, Sagemaster, AI Data Defense
- **Sivilsamfunn og uavhengige aktører:** Teknologirådet, Abelia, Green Producers Club, OUS, selvstendig næringsdrivende, frilansere

Dagens agenda

- 10:00–10:15 Kaffe og mingling
- 10:15–10:30 Velkomst og introduksjon – Sagar Sen & Gencer Erdogan, SINTEF Digital
- 10:30–11:00 Malcolm Langford (UiO): Juridiske perspektiver
- 11:00–11:30 Finn Myrstad (Forbrukerrådet): Retningslinjer for offentlig sektor
- 11:30–12:00 Andreas Hafver (DNV): Rammeverk for trygg bruk av KI
- 12:00–12:45 Lunsj
- 12:45–13:15 Milan De Cauwer (SINTEF): Feilklassifiseringsdetektorer
- 13:15–13:45 Boning Feng (OsloMet): KI og cybersikkerhet
- 13:45–14:00 Oppsummering og avslutning.