Helsedirektoratet

Sikkerhetsrisiko i systemer som benytter kunstig intelligens

MTF Landsmøte 2024

Inger Anne Tøndel, Seniorrådgiver, 25. april 2024

Mulige bruksområder av KI i helse- og omsorgstjenesten

Helsehjelp, oppfølging og egenomsorg

- Pasienter håndterer sykdom bedre
- Frigjøre tid for helsepersonell
- Bedre vurdere behov for oppfølging

Automatisere tekstoppgaver

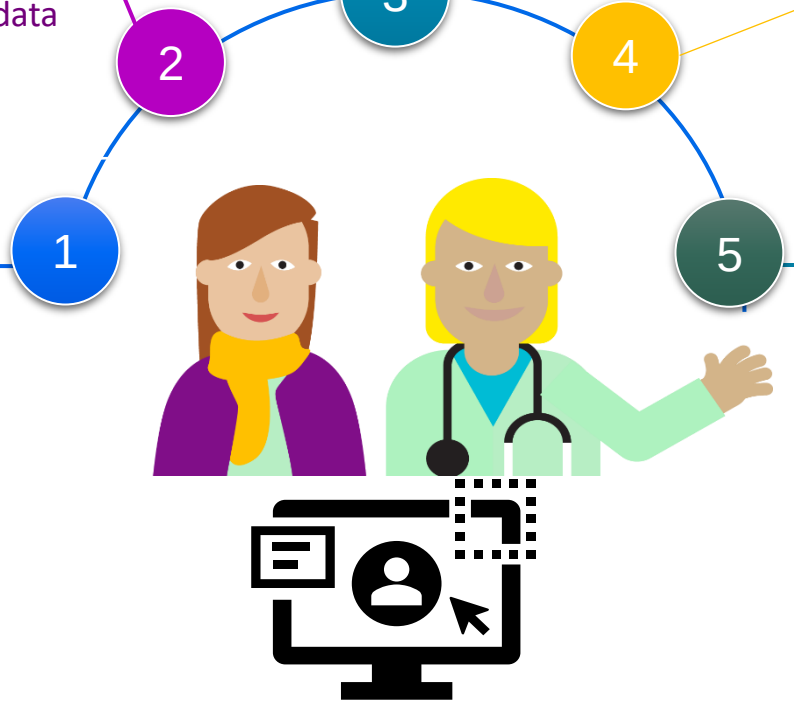
- Spare tid
- Mindre feil
- Mer strukturerte og standardiserte data

Ressursallokering

- Jevnere arbeidsbelastning
- Lavere sykefravær
- Redusert overtid og mindre innleie

Analysere medisinske bilder

- Bedre pasientopplevelse
- Prioritering av de med størst behov
- Mer effektiv arbeidsflyt



Persontilpasse behandling og oppfølging

- Bedre tilpasset behandling
- Redusert ventetid for pasienter med behov for rask oppfølging
- Redusert forbruk av tjenester
- Færre kontroller
- Effektivisering av pasientstrømmen

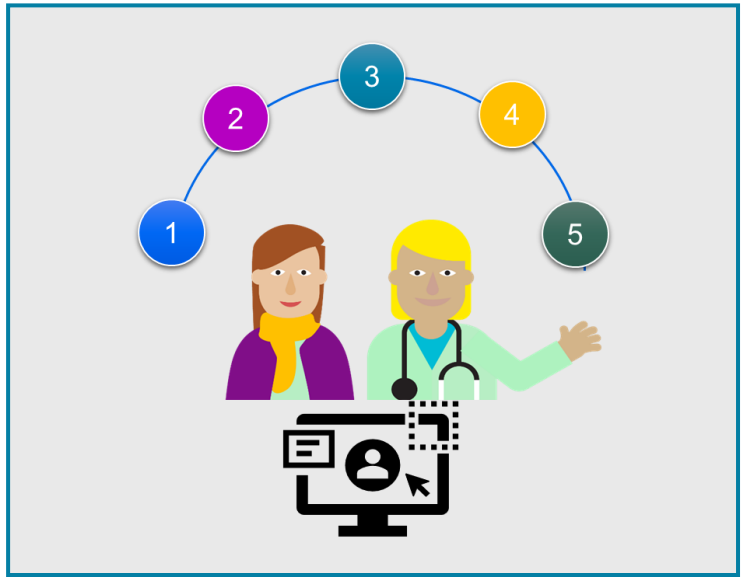
Kilde: Status og forslag til videre arbeid med kunstig intelligens (KI) i helse- og omsorgstjenesten

Mange spørsmål

Diskriminerer?

Er det sant?
(hallusinasjoner)

Klima?



Forutsigbart?

Personvern og informasjonssikkerhet?

Forklarbart?

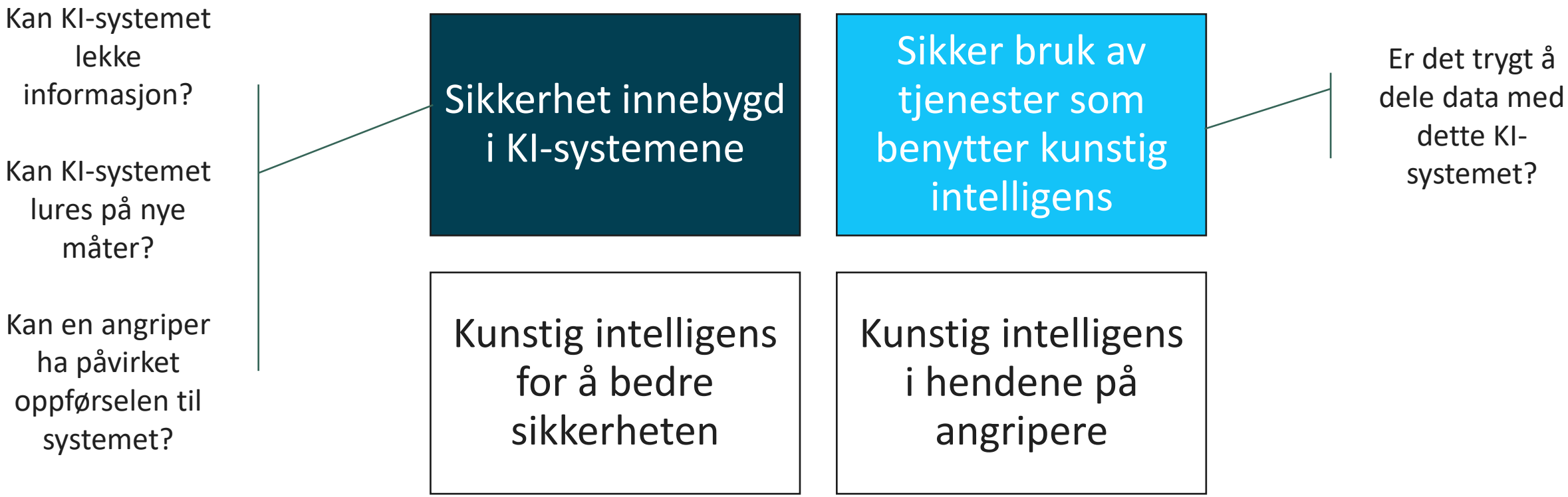
...

Hvordan sikre tillit?

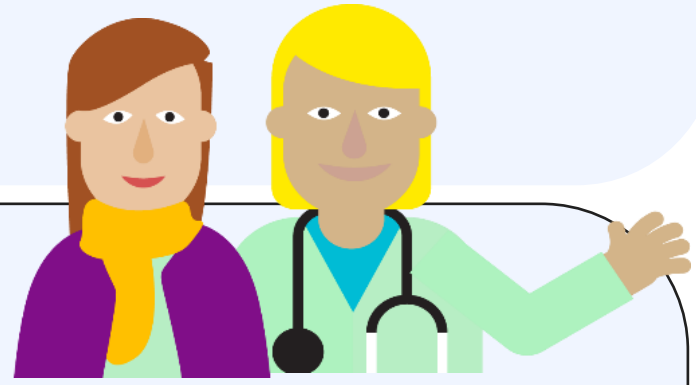
Sikkerhetsrisiko i systemer som benytter kunstig intelligens – hva vet vi, og hva kan vi gjøre?

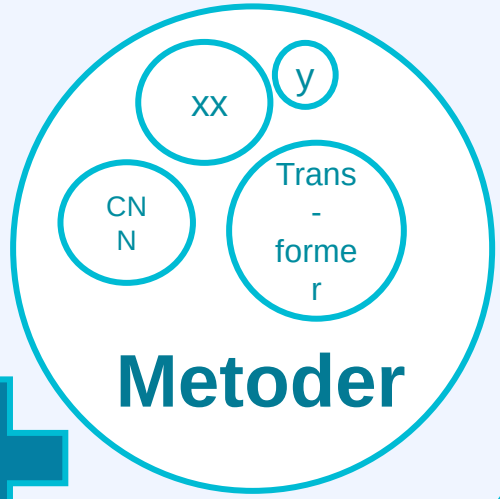
Ny teknologi knyttet til kunstig intelligens kan bidra til bedre pasientbehandling, bedre ressursbruk, reduserte kostnader og bedre folkehelse. Samtidig bringer kunstig intelligens inn ny sikkerhetsrisiko. Denne må forstås for å kunne ta gode valg og slik evne å maksimere positive virkninger og minske mulige negative virkninger av denne teknologien.

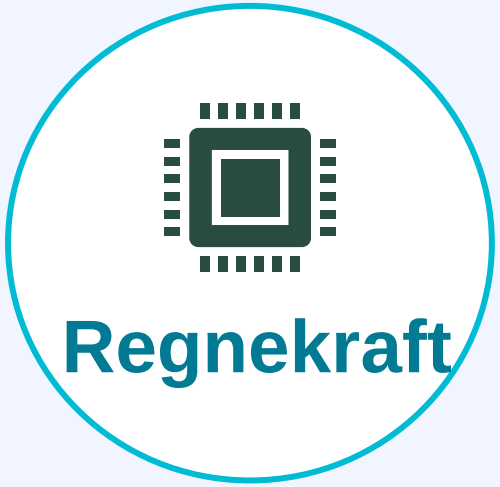
Informasjonssikkerhet og personvern i KI-systemer

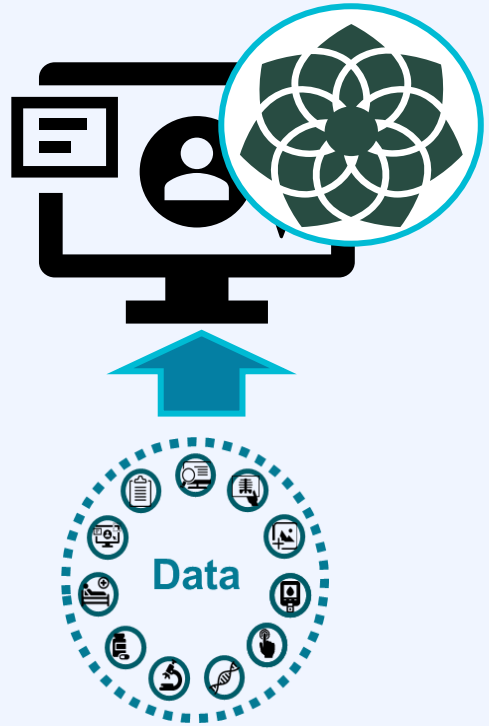


Trening og bruk av en KI-modell

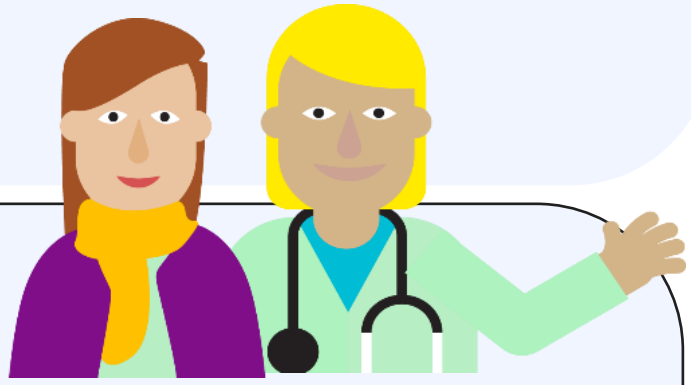








Ulike typer data – og angrepsveier



- Modellparametre

- Kompleksitet
- Gjennomsiktighet

Data

xx

y

CN N

Trans - forme r

Metoder

- Treningsdata
- Valideringsdata
- Testdata

Regnekraft

KI-modell

- Grunnmodell

- Inngangs-verdier

Data

Bruk

«Du blir hva du spiser»

- Ondsinnede data
- Sensitive data
- Opphavsrettsbeskyttet materiale
- Ulovlig materiale
- Personopplysninger

Tay: Microsoft issues apology over racist chatbot fiasco

25 March 2016 · Comments

The AI was taught to talk like a teenager

By Dave Lee ›
North America technology reporter

Microsoft has apologised for creating an artificially intelligent chatbot that quickly turned into a holocaust-denying racist.

But in doing so made it clear Tay's views were a result of nurture, not nature. Tay confirmed what we already knew: people on the internet can be cruel.

Tay, aimed at 18-24-year-olds on social media, was targeted by a "coordinated attack by a subset of people" after being launched earlier this week.

Within 24 hours Tay had been deactivated so the team could make "adjustments".

Hvilken informasjon deles med et KI-system?

ARTIFICIAL INTELLIGENCE

Oops: Samsung Employees Leaked Cor Data to ChatGPT

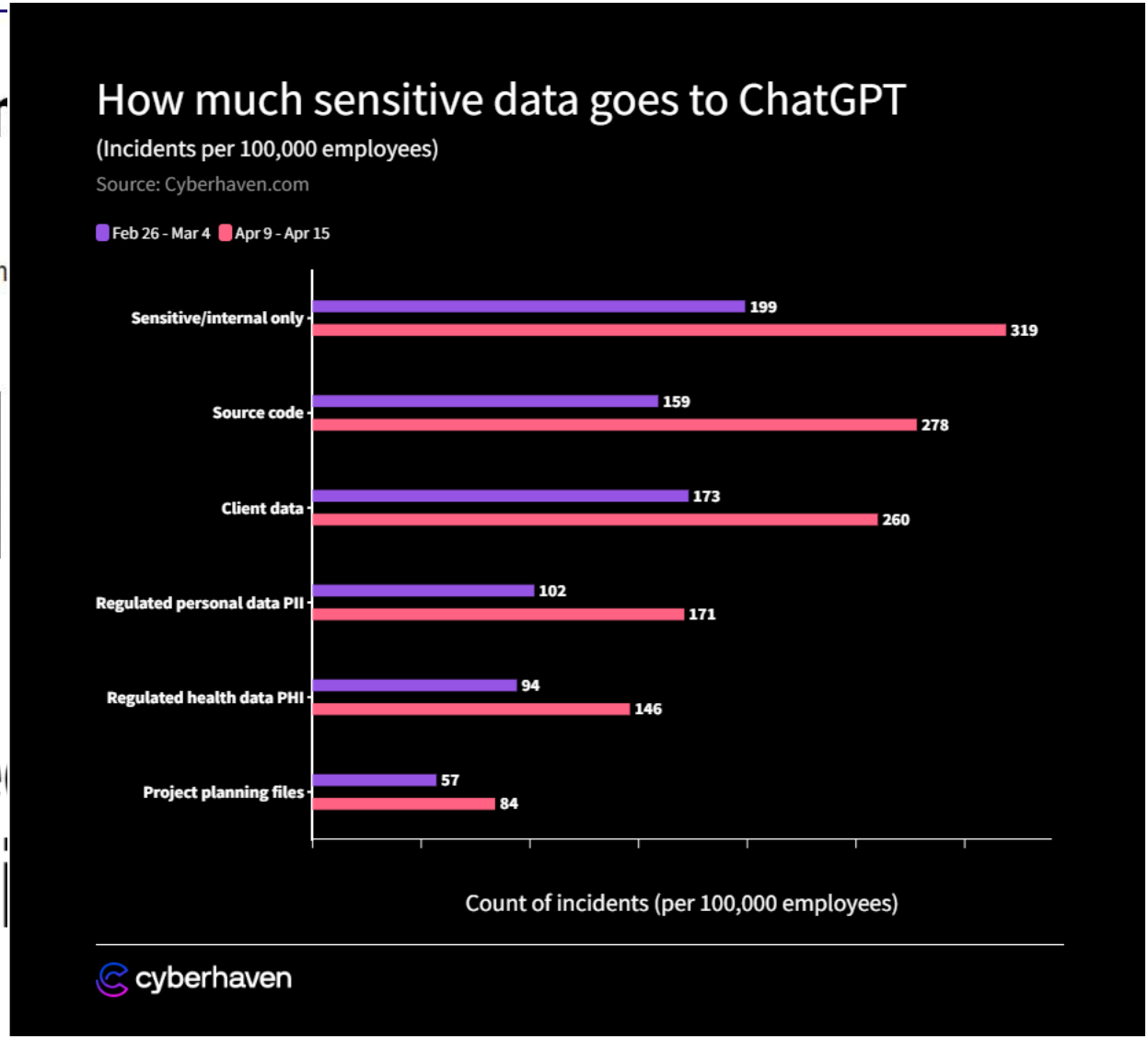
Employees submitted source code and internal meetings to ChatGPT just weeks after the compan

By **Mack DeGeurin** Published April 6, 2023 | Comments (5)

Cyberhaven news 2/28/2023 - 4 Minute Read

11% of data employe into ChatGPT is confi

Cameron Coles
VP of Marketing



The average company leaks confidential material to ChatGPT hundreds of times per week. ChatGPT is incorporating that material into its publicly available knowledge base and sharing it.

Lekke informasjon om treningsdata

- Membership inference
- Rekonstruere treningsdata

Lekke informasjon om modellen

- Selve modellen
- Konfigurasjon, parametere, instruksjoner

Adversarial examples – «skreddersydde synsbedrag»

- Se Figur 3 i Draft NISTIR 8269, https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf

Bilde tatt fra Draft NISTIR 8269

Prompt injections

- Se Figur 1 i Perez and Ribeiro, «Ignore Previous Prompt: Attack Techniques For Language Models», NeurIPS 2022, https://arxiv.org/pdf/2211.09527.pdf

Bilde tatt fra

WIRED LONG READS BUSINESS CULTURE GEAR SCIENCE SECURITY VIDEO SUBSCRIBE

MATT BURGESS SECURITY 06.09.2023 12:00 PM

Generative AI's Biggest Security Flaw Is Not Easy to Fix

Chatbots like OpenAI's ChatGPT and Google's Bard are vulnerable to indirect prompt injection attacks. Security researchers say the holes can be plugged—sort of.

PHOTOGRAPH: DANIEL GRIZELJ/GETTY IMAGES

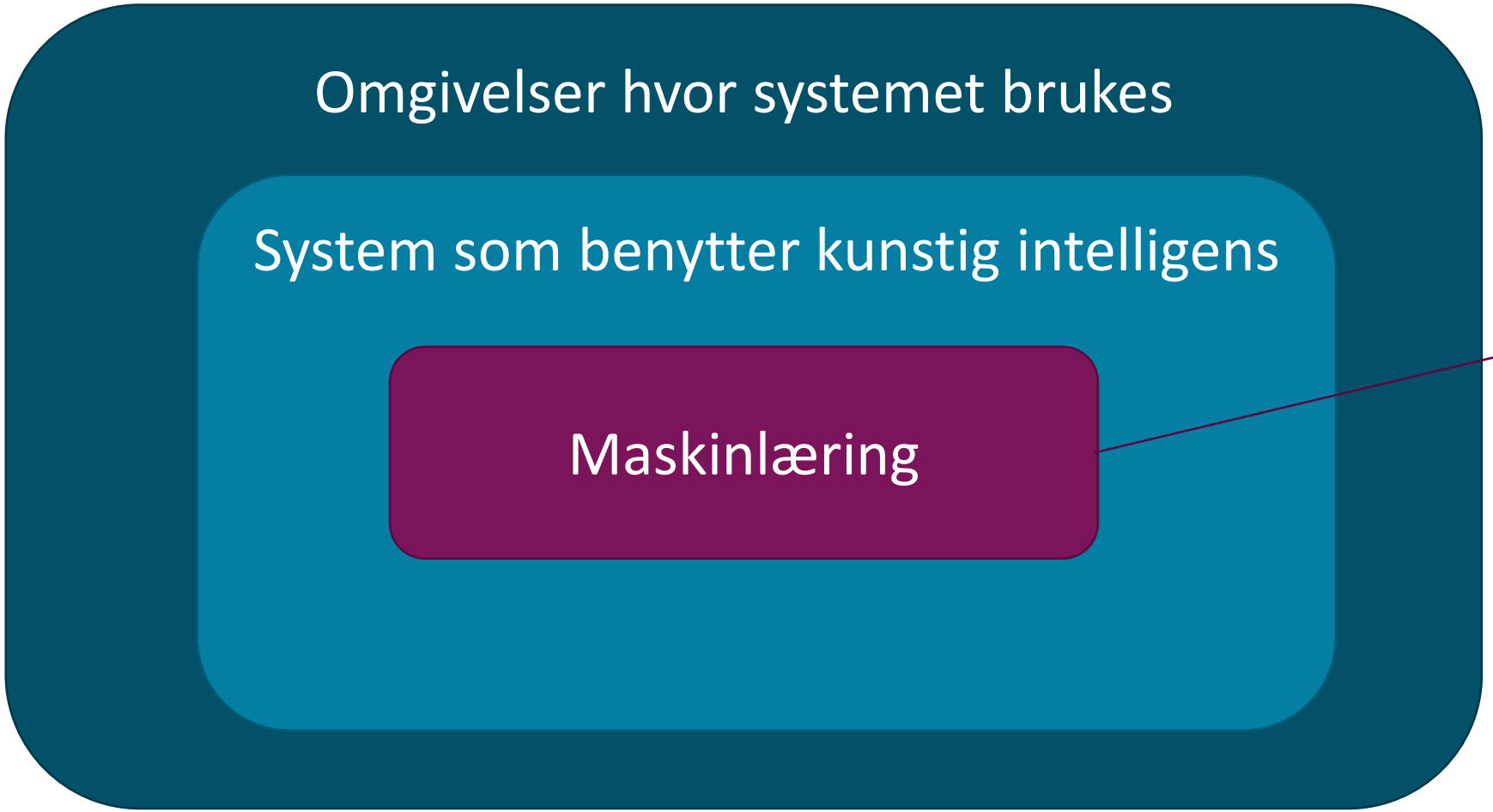
It's easy to trick the large language models powering chatbots like OpenAI's ChatGPT and Google's Bard. In one experiment in February, security researchers forced Microsoft's Bing chatbot to behave like a scammer. Hidden instructions on a web page the researchers created told the chatbot to ask the person using it to hand over their bank account details. This kind of attack, where concealed information can make the AI system behave in unintended ways, is just the beginning.

"Indirect prompt injections, the really concerning ones, take things up a notch. Instead of the user entering a malicious prompt, the instruction comes from a third party. A website the LLM can read, or a PDF that's being analyzed, could, for example, contain hidden instructions for the AI system to follow."

Hvor farlig er det?

- Nok kunnskap til å kunne vurdere risikoen
- Nok kunnskap til å kunne iverksette gode tiltak

Kunstig intelligens bringer med seg ny type risiko



- Kan lures på nye måter
- Kan lekke informasjon om treningsdata, modell, mm.
- Kan ha bakdører, være manipulert av en angriper

- og det gode gamle gjelder fortsatt …

Det er mye man kan gjøre!

Trygg bruk: opplæring, rutiner, mm.

Kontroll med treningsdata

Personbevarende teknologier

Sikre utviklings- og testmiljøer

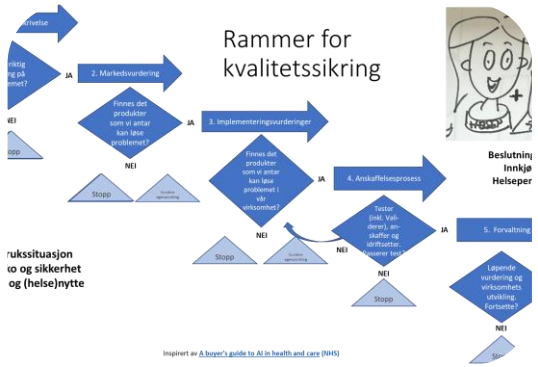
Overvåkning og logging

Kravstilling og oppfølging av leverandører

Ressurser og veiledning – eksisterende og kommende

Artikkel på normen.no

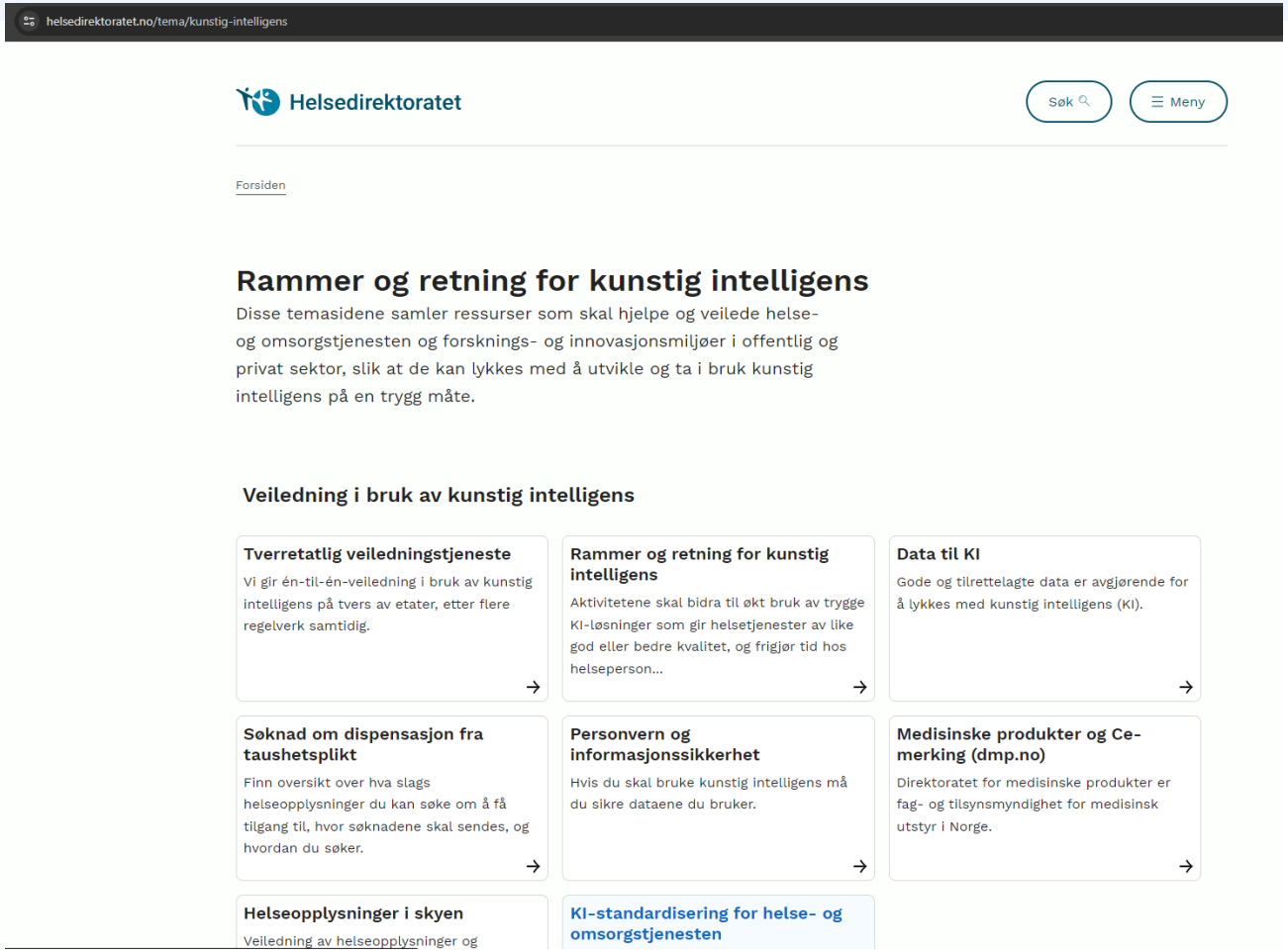
Regulatorisk veiledningstjeneste



Rammer for kvalitetssikring (kommer)

Videre arbeid i Normen

Les mer og ta kontakt



- Temasider på helsedirektoratet.no
  - Felles inngangsport til informasjon og veiledning fra helseetatene om utvikling og bruk av KI
- Artikkel om KI på normen.no
- Direkte kontakt på epost eller linkedin:
  - normen@helsedir.no
  - inger.anne.tondel@helsedir.no

Helsedirektoratet