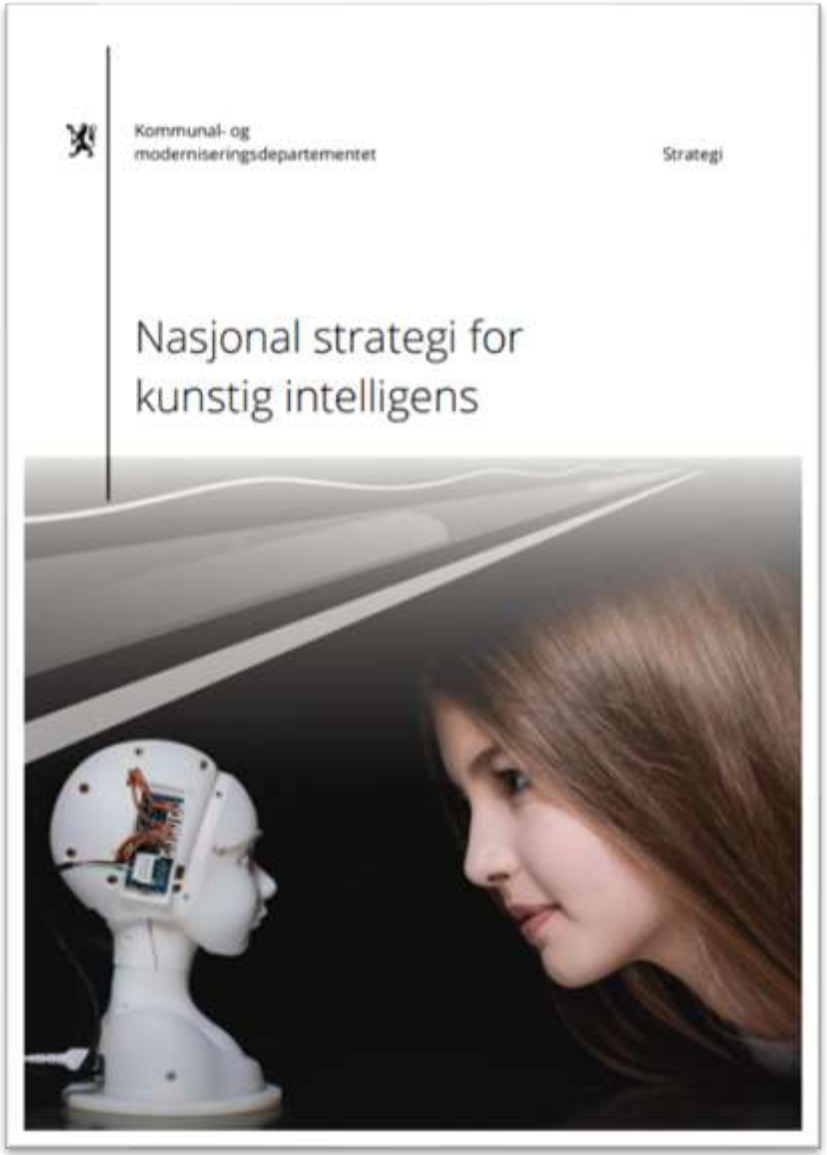
# NTNU/Copilot – Sandkasseprosjekt våren 2024

20. mars 2024 – «Løypemelding 1»

Eirik Gulbrandsen | Senioringeniør seksjon Teknologi, Sikkerhet og Tilsyn

Datatilsynet

# Nasjonal strategi for kunstig intelligens




*«Regjeringen vil at Norge skal gå foran i utvikling og bruk av kunstig intelligens med respekt for den enkeltes rettigheter og friheter.»*

**Tiltak for ansvarlig innovasjon:**

Regulatorisk sandkasse* for personvern og kunstig intelligens

**Utvidet mandat 2023:**

Regulatorisk sandkasse* for personvernvennlig innovasjon og digitalisering

** Regulatorisk sandkasse – «utvidet veiledning»*

# Hva er en regulatorisk sandkasse?

et kontrollert miljø for virksomheter som vil eksperimentere med nye produkter, teknologier og tjenester under oppfølging av Datatilsynet

**Virksom-heter**

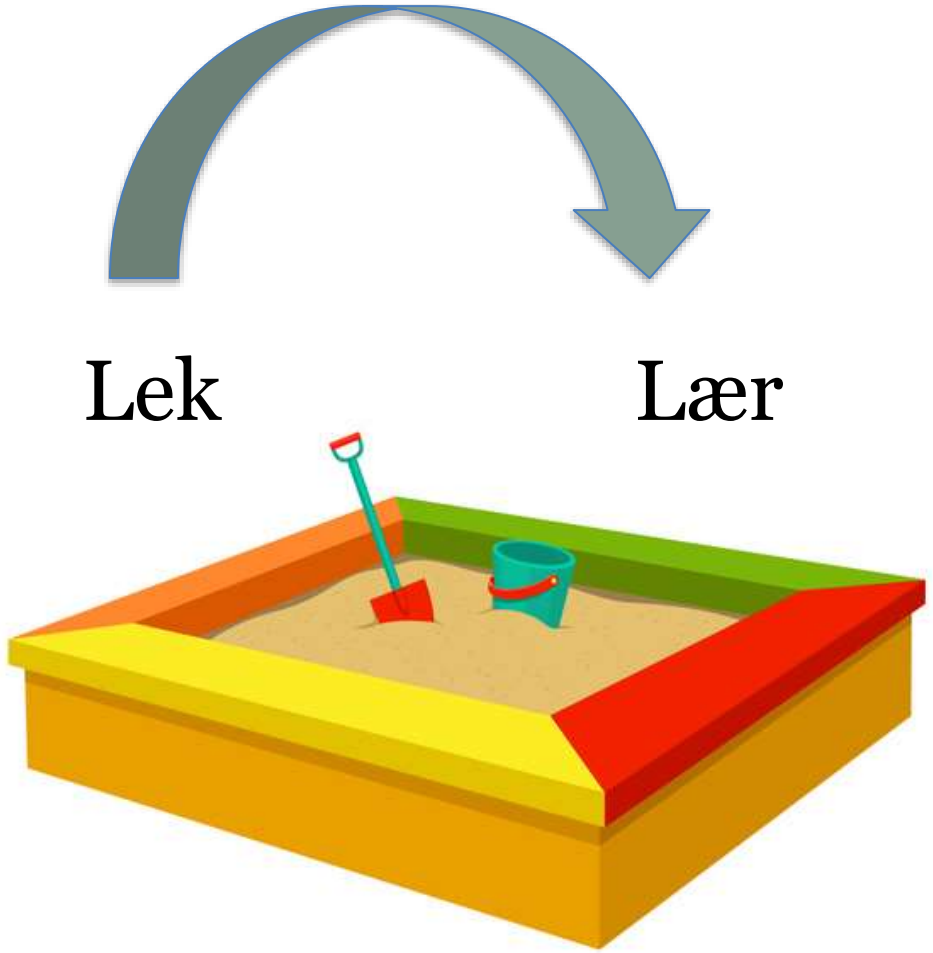
- **Bedre regelverksforståelse** – for de som deltar, og for andre virksomheter
- Bidra til **personvernvennlige løsninger**

**Datatilsynet**

- Øke tilsynets **kunnskap og forståelse av innovative løsninger.**
- Utarbeide **veiledning.**

**Enkeltindivider og samfunnet**

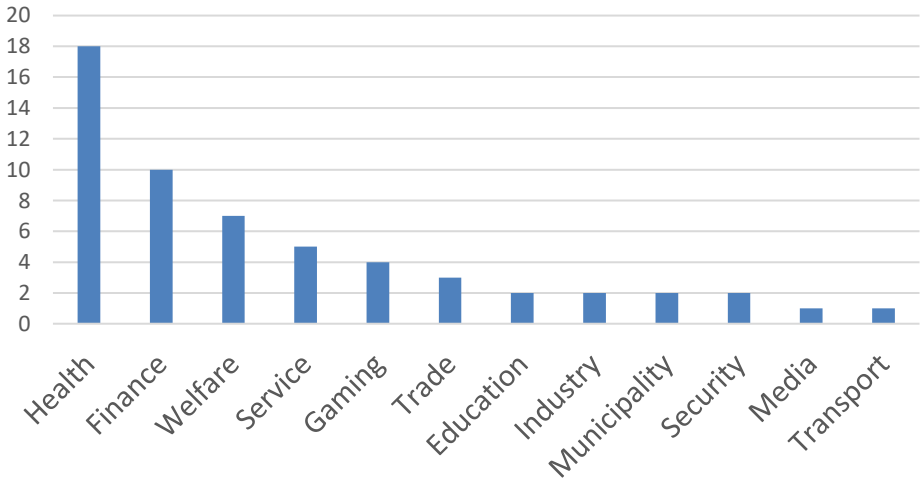
- **Bygge tillit** til ny bruk av innovativ løsninger innenfor **ansvarlige rammer**.



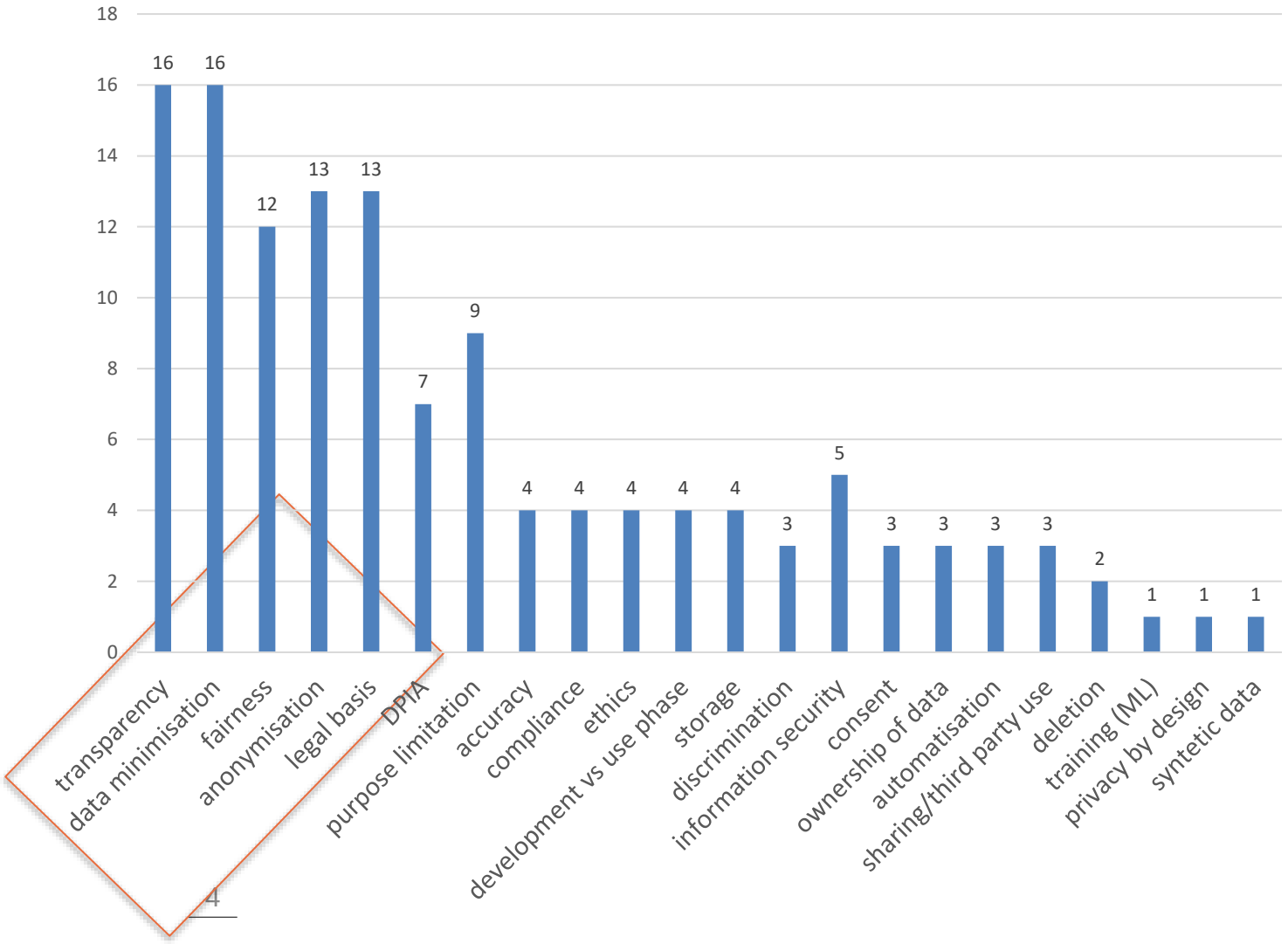
# Datatilsynets sandkasse: over 100 søkere – 12 prosjekter

## Fordeling offentlig / privat


## Søkere pr sektor




## Tema pr søker



# Datatilsynets sandkasseprosjekter våren 2024

1. **NTNU (stor/offentlig)**
   – Teste Microsoft KI-assistent «Microsoft 365 Copilot», med fokus på å vurdere verktøyets muligheter og utfordringer, samt etablering av et robust rammeverk for forvaltning, drift, vedlikehold og utvikling.

2. **Helsedirektoratet (stor/offentlig)**
   – Forenklet tilgang til informasjon gjennom generative språkmodeller.

3. **JuridiskABC (liten/privat)**
   – Bruk og videreutvikling av juridisk chatbot basert på KI/språkmodeller. Tjenesten er rettet mot virksomheter, særlig HR og ledere med fokus på arbeidsrettslige spørsmål.

# Uregulert kunstig intelligens?

## Eksempler på sektorovergripende lover:

- GDPR (personvern)
- Handelspraksisdirektivet (villedende/aggressiv handel)
- Produktsikkerhetsdirektivet (utrygge produkter)
- Forordning om digitale tjenester – DSA (innholdsmoderering)

## Eksempler på sektorspesifikke lover:

- Forvaltningsloven (rett til forklaring)
- Pasientjournalloven, helseregisterloven, helseforskningsloven
- Hvitvaskingsloven (krav til overvåkingssystemer)
- Finansforetaksloven (risikovurderinger)

**Utvalg nye rettsakter fra EU som er datarelevante:**

**Vedtatte:**
Critical Entities Resilience Directive (CER)
Data Act
Data Governance Act (DGA)
Digital Markets Act (DMA)
Digital Operational Resilience Act (DORA)
Digital Services Act (DSA)
General Product Safety Regulation (GPSR)
NIS 2 Directive

**Under utarbeidelse:**
AI Act
AI Liability Directive
Consumer Credit Directive
CSAM Regulation
Cyber Resilience Act
Cyber Security Act
Cyber Solidarity Act
ePrivacy Regulation
European Health Data Space (EHDS)
European Media Freedom Act
Platform Work Directive
Political Advertising Regulation
Product Liability Directive (PLD)

# KI-strategi / forslag til KI-regulering (Høyre)

STORTINGET

Representantforslag 46 S

(2023–2024)

fra stortingsrepresentantene Linda Hofstad Helleland, Nikolai Astrup, Anne Kristine Linnestad, Margret Hagerup, Sveinung Stensland og Svein Harberg

Dokument 8:46 S (2023-2024)

Representantforslag fra stortingsrepresentantene Linda Hofstad Helleland, Nikolai Astrup, Anne Kristine Linnestad, Margret Hagerup, Sveinung Stensland og Svein Harberg om Norges neste steg for kunstig intelligens

1.2 Regulering

Lovreguleringer, kontroll og tilsyn er ikke noe mål i seg selv, men noe et moderne samfunn benytter for å sikre at den enkelte innbyggers og virksomhets plikter og rettigheter er tydelige og forutsigbare, og at det finnes systemer for å ivareta disse. Regulering skal også sikre at noens frihet ikke går på bekostning av andres mulighe-

STORTINGET SAKER REPRESENTANTER OG KOMITEER HVA SKJER?

Du er her: Forsiden - Saker og publikasjoner - Finn saken - Sak

Representantforslag om kunnskap om og veiledning i bruk av kunstig intelligens

Dokument 8:273 S (2022-2023), Innst. 151 S (2023-2024) Følg saken (e-postvarsling)

STORTINGET SAKER REPRESENTANTER OG KOMITEER HVA SKJER?

Du er her: Forsiden - Saker og publikasjoner - Finn saken - Sak

Representantforslag om demokratisk kunstig intelligens

Dokument 8:232 S (2022-2023), Innst. 152 S (2023-2024) Følg saken (e-postvarsling)

**1.** Opprette et rådgivende og frittstående organ for å følge med på utviklingen av kunstig intelligens, gi råd i prinsipielle spørsmål rundt teknologien og veilede om ansvarlig bruk av den. Organet kan på sikt videreutvikles til å bli et lovhjemlet algoritmetilsyn.

**2.** Gjennomføre en digital lovvask, der det kartlegges hvilke eksisterende lov- og regelverk som er godt egnede rammeverk for teknologi som kunstig intelligens, og hvor det utarbeides lover der det er mangler. Lovvasken bør ha særlig fokus på åpenhet og innsyn, rettssikkerhet, personvern og forbrukerrettigheter

**3.** Styrke kompetansen og kapasiteten hos eksisterende tilsyn på personvern-, forbruker- og likestillings- og diskrimineringsområdet for å

# Uregulert kunstig intelligens?

## Eksempler på sektorovergripende lover:

- **GDPR (personvern)**
- Handelspraksisdirektivet (villedende/aggressiv handel)
- Produktsikkerhetsdirektivet (utrygge produkter)
- Forordning om digitale tjenester – DSA (innholdsmoderering)

## Eksempler på sektorspesifikke lover:

- Forvaltningsloven (rett til forklaring)
- Pasientjournalloven, helseregisterloven, helseforskningsloven
- Hvitvaskingsloven (krav til overvåkingssystemer)
- Finansforetaksloven (risikovurderinger)

**Utvalg nye rettsakter fra EU som er datarelevante:**

**Vedtatte:**
Critical Entities Resilience Directive (CER)
Data Act
Data Governance Act (DGA)
Digital Markets Act (DMA)
Digital Operational Resilience Act (DORA)
Digital Services Act (DSA)
General Product Safety Regulation (GPSR)
NIS 2 Directive

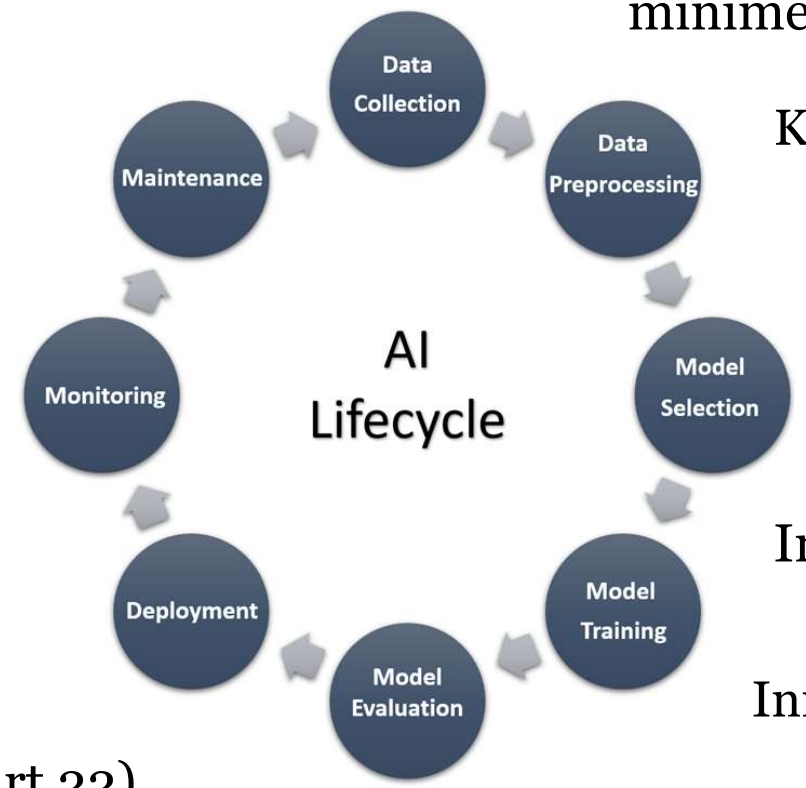
**Under utarbeidelse:**
AI Act *
AI Liability Directive
Consumer Credit Directive
CSAM Regulation
Cyber Resilience Act
Cyber Security Act
Cyber Solidarity Act
ePrivacy Regulation
European Health Data Space (EHDS)
European Media Freedom Act
Platform Work Directive
Political Advertising Regulation
Product Liability Directive (PLD)

* KI med uakseptabel risiko; måle ansattes følelser på arbeid eller studie

# GDPR regulerer KI

- Trening, validering og testing når datasettene inneholder personopplysninger
- Inneholder selve modellen personopplysninger?
- Inn-dataene (*prompts)* kan brukes til å forbedre modellen
- Ut-dataene kan si noe om enkeltpersoner
- KI kan brukes til å avdekke personopplysninger
- KI kan brukes til å foreta helautomatiserte avgjørelser

# Kunstig intelligens og personvernforordningen (GDPR)




Formål (art 5)

Data-minimering (art 5)

Korrekthet (art 5)

Lovlig (art 6)

Særlige kategorier(art 9)

Informasjon (art 12)

Innsyn (art 15)

DPIA (art 35)

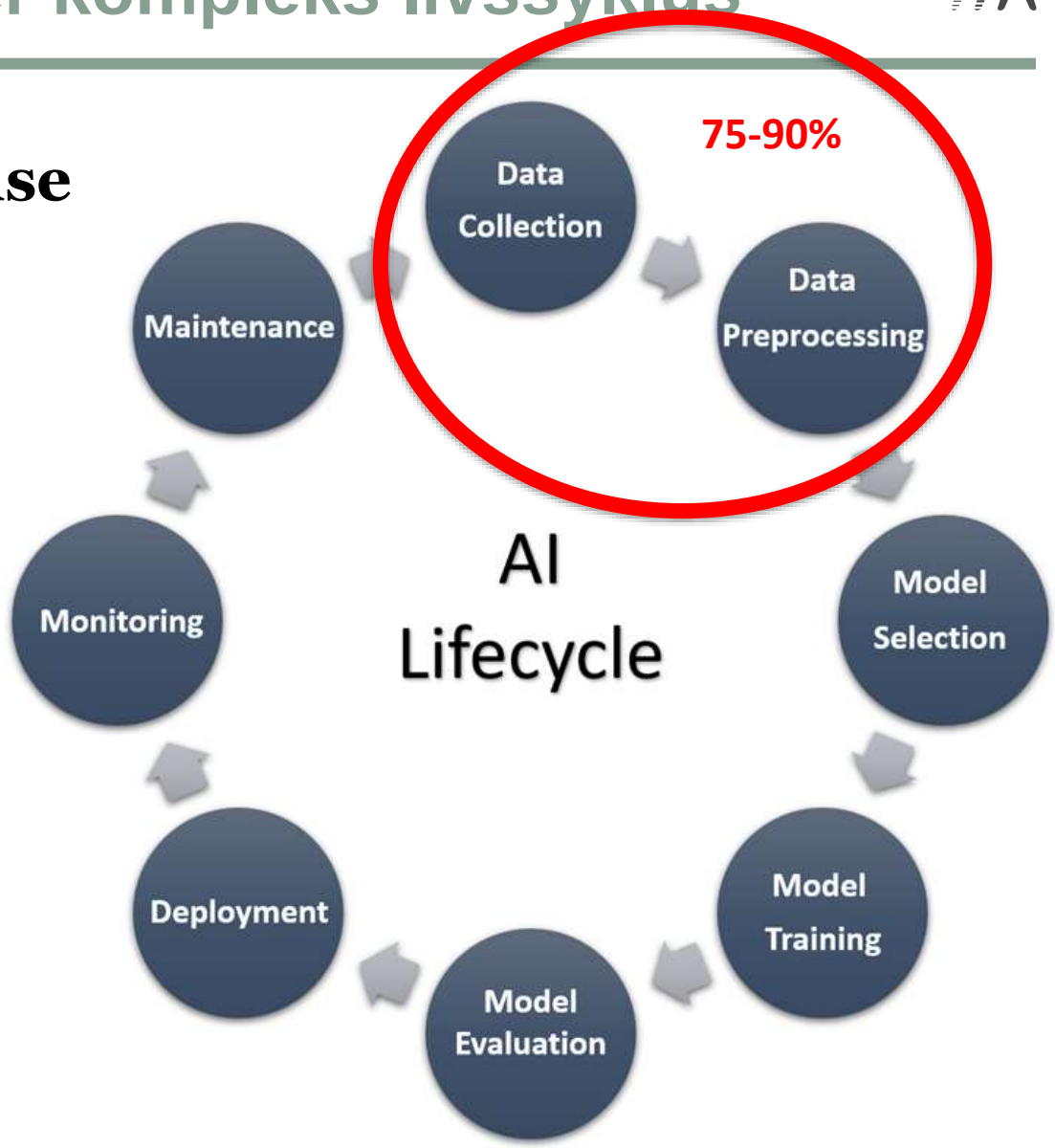
Innebygd personvern (art 25)

Automatisert behandling (art 22)

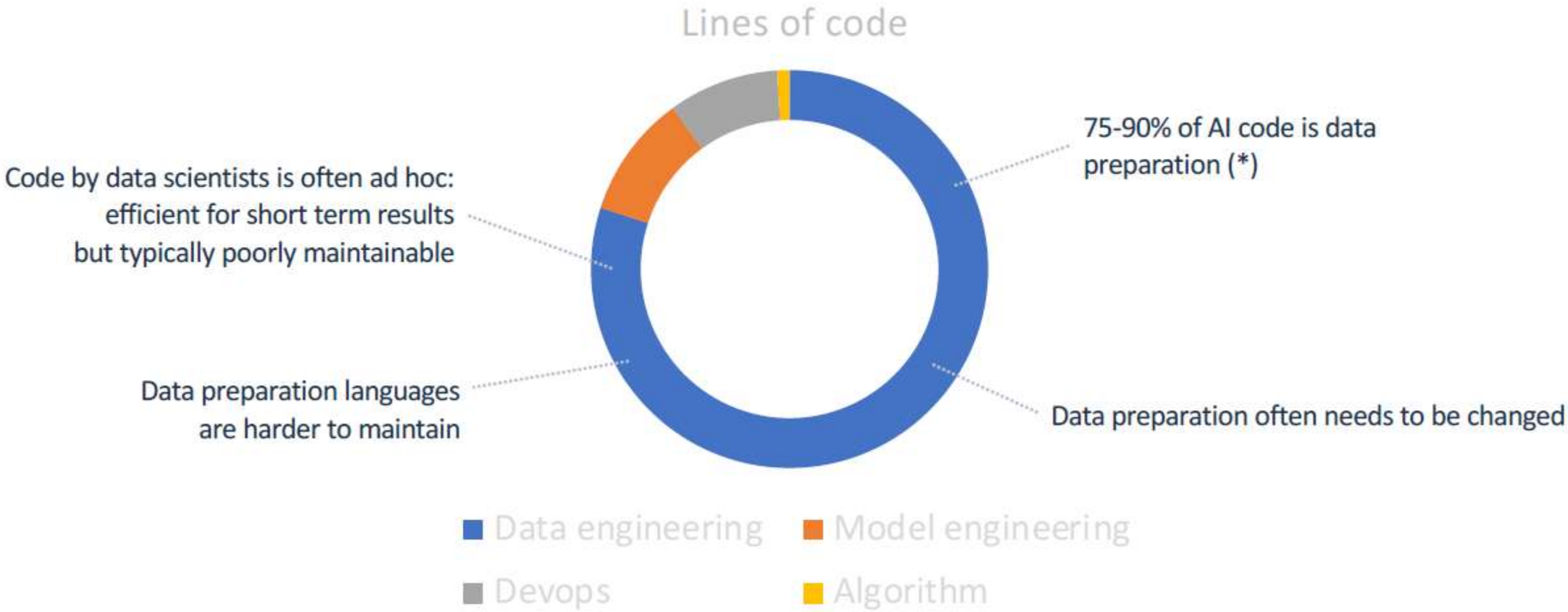
Sikkerhet ved behandlingen (art 32)

# Utfordringer spesifikke for KI – mer kompleks livssyklus

- **Krav til datavitenskap/dataforståelse**
- Mekanismer for å fange opp og håndtere skjevhet/ bias/ forklarbarhet/tolkbarhet
- Kontinuerlig oppfølging av kvalitet (eks. etterlæring)



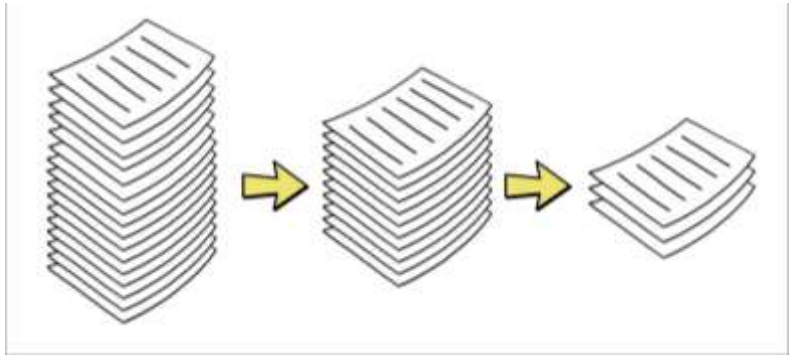
# Utfordringer spesifikke for KI – datapreparing eget fag



# 1: KI møter dataminimering

## KI

- Trenger mye data – og du vet ikke alltid akkurat hva du trenger



**VS.**

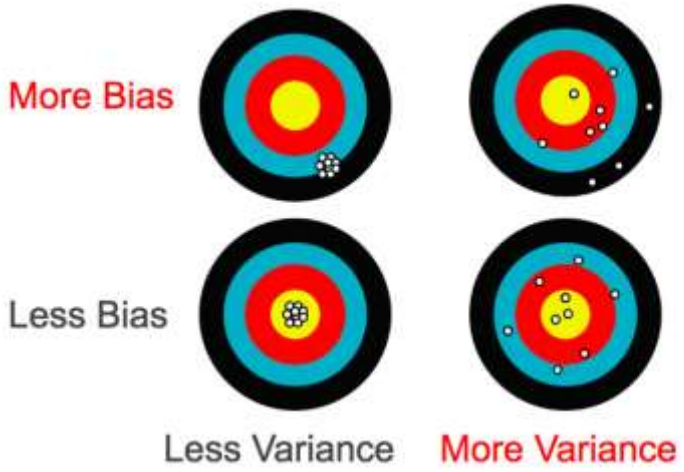
## Dataminimering

- Personopplysninger skal være adekvate, relevante og begrenset til det som er nødvendig for formålene de behandles for (art. 5)
- Formålsbegrensing: Personopplysninger skal samles inn for spesifikke, uttrykkelig angitte og berettigede formål

# 2: Skjeve algoritmer møter retten til rettferdighetsprinsippet

## Skjeve algoritmer, datakvalitet

- Shit in = shit out




**VS.**

## Rettferdighetsprinsippet

- Personopplysninger skal behandles på en lovlig, rettferdig og åpen måte med hensyn til den registrerte (art. 5)

# 3: Den svarte boksen møter krav til åpenhet og forklarbarhet

## Den svarte boksen

- Hva skjer inni der?


**VS.**

## Åpenhet / forklarbarhet

- Rett til innsyn, generell info + forklar logikken (art. 13, 14 & **15**)
- Retten til en forklaring (art. 22)
- Klart og forståelig språk (art. 12)
- Åpenhet (art. 5)

**GDPR artikkel 15 - Den registrertes rett til innsyn**

…relevant informasjon om den underliggende logikken samt om betydningen og de forventede konsekvensene…

# Sandkasseprosjekt: NAV

## NAV - sluttrapport

Våren 2021 startet sandkasseprosjektet som tar for seg NAVs KI-verktøy for å predikere utviklingen av sykefravær. Prosjektet ble avsluttet høsten 2021. Her er sluttrapporten fra prosjektet.

**Innhold**

1. Sammendrag
2. Om prosjektet
3. Rettslig grunnlag
4. Rettferdighet
5. Hvordan forklare bruken av kunstig intelligens?
6. Veien videre

Skriv ut alt innholdet

Last ned PDF

Søk i dette innholdet

## Sammendrag

NAV ønsker å bruke maskinlæring til å forutse hvilke sykmeldte brukere som vil ha behov for oppfølging to måneder frem i tid. Dette skal hjelpe veilederne med gjøre mer treffsikre vurderinger, som igjen skal spare NAV, arbeidsgivere og de sykmeldte for unødvendige møter. Målet med dette sandkasseprosjektet var å avklare lovligheten ved bruk av kunstig intelligens (KI) i denne sammenhengen, og utforske hvordan profileringen av sykmeldte kan gjøres på en rettferdig og åpen måte.

### Konklusjoner

1. **Lovlighet.** NAV har rettslig grunnlag for å bruke KI som støtte ved beslutning om enkeltindividers behov for oppfølging og dialogmøte. Det er usikkert om det rettslige grunnlaget åpner for å bruke personopplysninger til å utvikle selve algoritmen.
2. **Rettferdighet.** Det er viktig forskjell mellom å benytte opplysninger som allerede inngår i modellen, og å ta i bruk nye opplysninger som ikke brukes i modellen, til å sjekke for diskriminerende utfall. Det oppstår en spenning mellom personvern og rettferdighet når metoden for å avdekke og motvirke diskriminering fordrer mer *behandling av personopplysninger*.
3. **Åpenhet.** For at modellen skal gi ønsket verdi, er det avgjørende at NAV-

- *«Catch-22»*
  - *Hjemmelsgrunnlag for å bruke maskinlæringsmodeller*
  - *Ikke hjemmelsgrunnlag for å trene maskinlæringsmodeller*
- *Grunnlag for lovarbeid og mulig hjemmel for behandling*
- Rettferdighet, åpenhet og forklarbarhet
  - For borgere
  - For saksbehandlere
  - For utviklere
  - For organisasjonen
  - For samfunnet / myndigheter

# Sandkasseprosjekt: NAV → Forklarbarhet og åpenhet

## NAV - sluttrapport

Våren 2021 startet sandkasseprosjektet som tar for seg NAVs KI-verktøy for å predikere utviklingen av sykefravær. Prosjektet ble avsluttet høsten 2021. Her er sluttrapporten fra prosjektet.

**Innhold**



Skriv ut alt innholdet

Last ned PDF

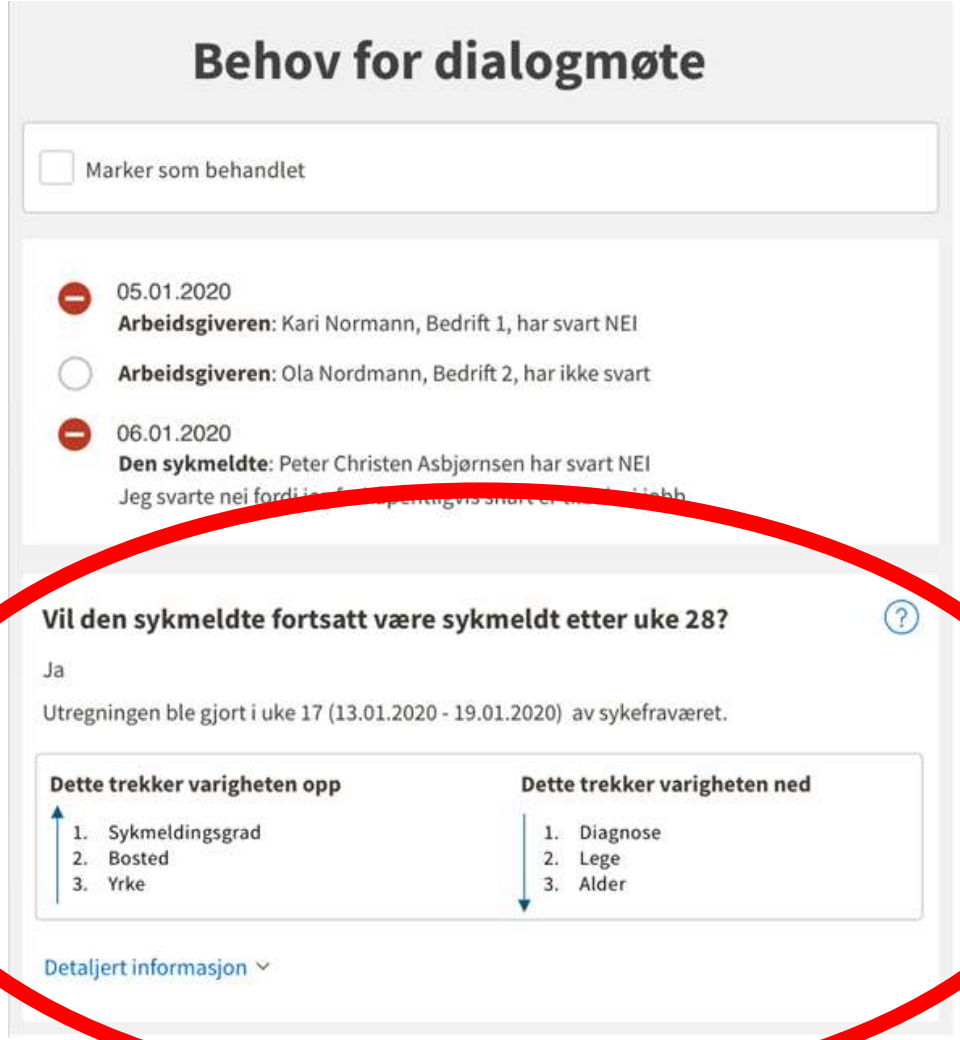
Søk i dette innholdet

## Sammendrag

NAV ønsker å bruke maskinlæring til å forutse hvilke sykmeldte brukere som vil ha behov for oppfølging to måneder frem i tid. Dette skal hjelpe veilederne med gjøre mer treffsikre vurderinger, som igjen skal spare NAV, arbeidsgivere og de sykmeldte for unødvendige møter. Målet med dette sandkasseprosjektet var å avklare lovligheten ved bruk av kunstig intelligens (KI) i denne sammenhengen, og utforske hvordan profileringen av sykmeldte kan gjøres på en rettferdig og åpen måte.

### Konklusjoner

1. **Lovlighet.** NAV har rettslig grunnlag for å bruke KI som støtte ved beslutning om enkeltindividers behov for oppfølging og dialogmøte. Det er usikkert om det rettslige grunnlaget åpner for å bruke personopplysninger til å utvikle selve algoritmen.
2. **Rettferdighet.** Det er viktig forskjell mellom å benytte opplysninger som allerede inngår i modellen, og å ta i bruk nye opplysninger som ikke brukes i modellen, til å sjekke for diskriminerende utfall. Det oppstår en spenning mellom personvern og rettferdighet når metoden for å avdekke og motvirke diskriminering fordrer mer *behandling av personopplysninger*.
3. **Åpenhet.** For at modellen skal gi ønsket verdi, er det avgjørende at NAV-



# Sandkasseprosjekt: Secure Practice

## Skreddarsydd kunnskap i kampen mot cyberkrim

Den andre sluttrapporten frå Datatilsynets sandkasse for kunstig intelligens slår to fluger i eitt smekk, og kan vere eit viktig steg på vegen både for betre beredskap mot cyberkriminalitet og for betre personvern i arbeidslivet.


### Personvernvenleg profilering

Secure Practice er eit norsk firma som tilbyr tenester for *informasjonstrygglеik*. I fjor var dei med i den regulatoriske sandkassa for ansvarleg kunstig intelligens, der dei saman med Datatilsynet utforska ei ny teneste firmaet utviklar og ønsker å få på marknaden. No er sluttrapporten frå prosjektet klar.

Kjernen i tenesta Secure Practice vil tilby, byggjer på erkjenninga av at menneskelege feil ofte er medverkande når hackerane lukkast. Å gi tilsette god opplæring i trygg passordhandtering, teikn på phishing og andre cybertruslar reduserer faren for hacking. Og jo meir skreddarsydd denne opplæringa er etter kunnskapsnivået, nettvanane og motivasjonen for kvar enkelt tilsett, jo meir effektiv vil den vere. Det er berre ein hake ved det. All informasjonen som trengs for å vite kva som fungerer på akkurat deg – kven skal ha tilgang til den? Er det muleg å få til

- Ny teknologi (maskinlæring)
  → Personvern-konsekvensvurdering (DPIA)

- Behandlingsgrunnlag
  - Arbeidsmiljøloven
  - E-postforskriften
  - Personvernforordningen
    - 6.1.f – berettighet interesse

- Felles behandlingsansvar
  - Tjenesteyter beholder kontroll med deler av grunnlagsdata (personopplysninger)

# Sandkasseprosjekt: Secure Practice


COMPUTERWORLD

NYHETSBREV | KJØP ABONNEMENT

MACWORLD | TELECOM REVY | IT-BRANSJEN | EVENT | WHITEPAPER | STILLING LEDIG | BUSINESS CONTENT | E-AVIS

## EU satser på trøndersk cybersikkerhet

Trønderbedriften Secure Practice skal styrke cybersikkerheten over hele Europa. 1 million EU-borgere får norsk sikkerhetsteknologi gjennom EUs satsning på digital omstilling. Oppdraget er verdt 29 millioner kroner.

**Sitat CEO Erlend Andreas Gjære:**

*«Da kan det jo også være gøy å vite/fortelle at takket være sandkassa så kan vi nå rulle ut norsk innovasjon til hele Europa.»*

Secure Practice scoret full pott på vurderingen av prosjektets relevans og samfunnsnytten for EU. Utslagsgivende var også nybrottsarbeidet deres i krysningspunktet mellom kunstig intelligens (KI), cybersikkerhet og personvern, hvor de har behandlet viktige spørsmål sammen med Datatilsynet i den regulatoriske sandkassen for ansvarlig KI.

# Sandkasseprosjekt: Secure Practice


COMPUTERWORLD

NYHETSBREV | KJØP ABONNEMENT

MACWORLD | TELECOM REVY | IT-BRANSJEN | EVENT | WHITEPAPER | STILLING LEDIG | BUSINESS CONTENT | E-AVIS

## EU satser på trøndersk cybersikkerhet

Trønderbedriften Secure Practice skal styrke cybersikkerheten over hele Europa. 1 million EU-borg får norsk sikkerhetsteknologi gjennom EUs satsning på digital omstillin Oppdra er verd 9 millioner kr .

**Sitat CEO Erlend as Gjær**

*«Da kan det jo og å være y å vite/fortelle at ta et være sandkassa så kan nå rulle t norsk innovasjon t hele Eu pa.»*

Til litt!

ecure Pra scoret fu pott på vurderingen v pro ektets elevans og amfunnsnytten for U. Uts gsgiv de var også nybrottsarbeidet res i ysning unktet m lom kunstig i elligen (KI), cybersikkerhet og personvern, hvor de har behandlet viktige spørsmål sammen med Datatilsynet i den regulatoriske sandkassen for ansvarlig KI.

# Algoritmer, tillit og sandkasser

Personvernbloggen
Datatilsynets blogg om personvernspørsmål

Om

## Algoritmer, tillit og sandkasser

av Eirik Gulbrandsen | jan 6, 2021

Mot slutten av 1800-tallet i USA undersøkte kjemikeren Harvey Wiley innholdet i en rekke industrielt produserte matvarer. 90 prosent av honningkrukkene inneholdt ikke honning. Lønnesirup var stort sett ikke lønnesirup. Og syltetøy bestod, som de nevnte matvarene, også stort sett av billig maissirup, hvor «syltetøy» var tilsatt epleskall for tekstur. Verre var det at melk ble tilsatt formaldehyd og gipspulver for å fremstå som frisk, hvit og ubedervet. Barn ble syke og døde. I 1906 opprettet USA sin første forbrukervernlov for å sikre trygg mat.[1]

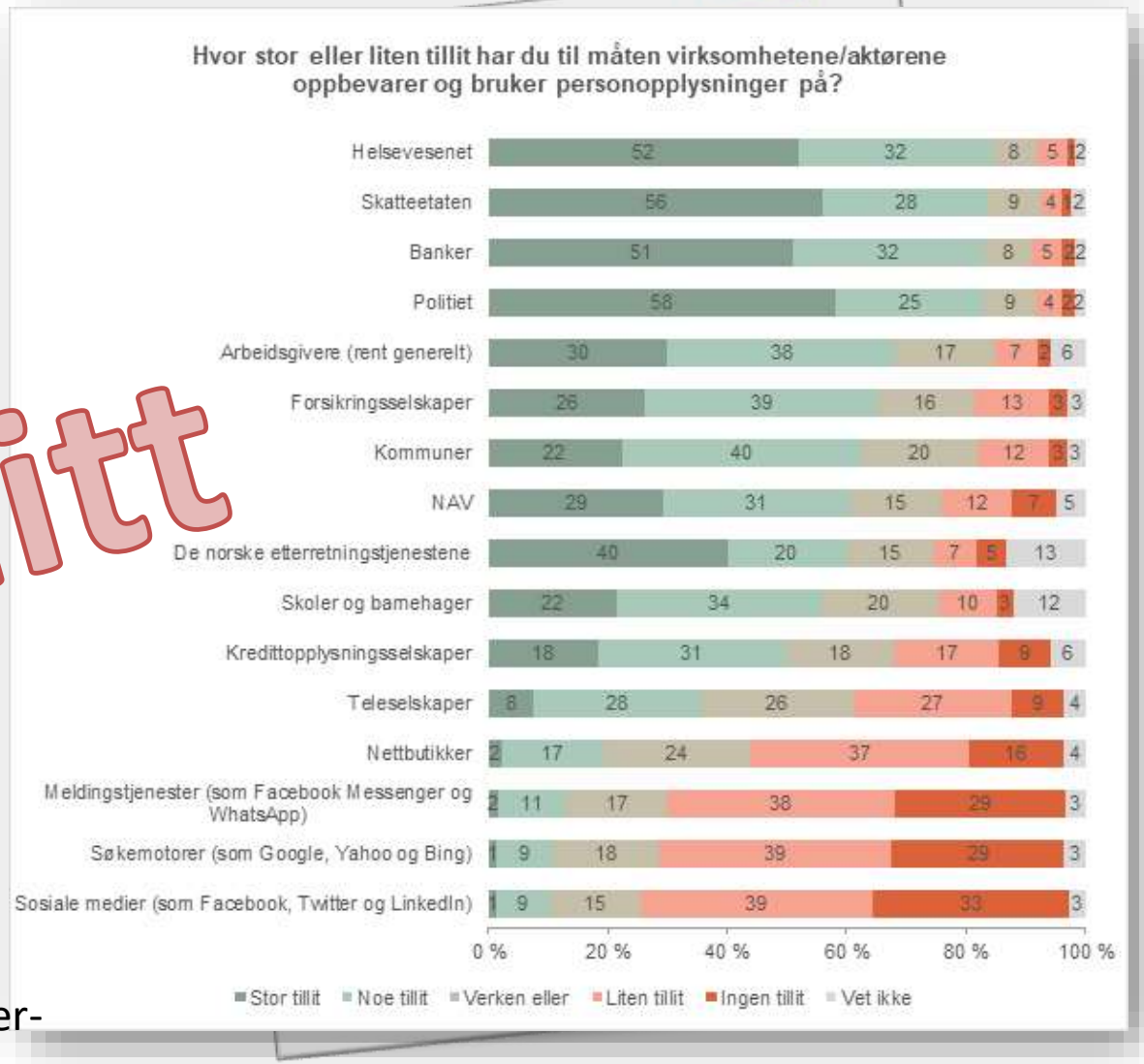
### En digital nedkjølingseffekt

Datatilsynet gjennomførte i årskiftet 2019/2020 en personvernundersøkelse som blant annet påviser en digital ned... Det betyr at mange begrenser bruken av digitale tjenester eller lar være å bruke dem. Effekten av digitalisering betyr... man ikke har den nødvendige tilliten til tjenestene – med rette.

Det blir stadig tydeligere at gigantene bak «sosiale medier» – som 1800-tallets matprodusenter – ikke tilbyr tjenestene til samfunnets og enkeltmenneskenes beste. Google, Facebook mfl. er i realiteten først og fremst annonsesalgsselskaper. Gjennom kynisk bruk av algoritmer, ønsker de først og fremst at du trykker på så mange annonser som mulig via deres plattformer.

> Dette er den moderne varianten av maissirup og utvannet melk – en dystopisk og polariserende digital virkelighet skapt fordi man ønsker å selge annonser.
>
> EIRIK GULBRANDSEN

Tillitt




https://www.personvernbloggen.no/2021/01/06/algoritmer-tillit-og-sandkasser/

# Forventninger til NTNU/Copilot prosjektet

- ✓ **Kunnskap om praktiske anvendelse av «generativ KI»**
- ✓ **Kunnskap om store aktørers «integrert KI»-løsninger**
- ✓ **Bidra til å identifisere og strukturere personvern-problemstillinger**
- ✓ **Bidra til å finne gode regulatoriske og praktiske tilnærminger til ansvarlig bruk = tillitt**
- ✓ **Bidra til å utvikle gode råd, veiledning og verktøykasse**

# Eirik Gulbrandsen

Datatilsynets sandkasse
Datatilsynets seksjon for Teknologi, Sikkerhet og Tilsyn

eirik.gulbrandsen@datatilsynet.no

Datatilsynet

postkasse@datatilsynet.no
Telefon: +47 22 39 69 00

**datatilsynet.no**
**personvernbloggen.no**