KI, samfunnssikkerhet og personvern

NIFS-møte 6. september: Kunstig intelligens og sikkerhet

Eirik Gulbrandsen | Senioringeniør seksjon Teknologi, Sikkerhet og Tilsyn

Datatilsynet

Algoritmer, tillit og sandkasser

Personvernbloggen

Datatilsynets blogg om personvernspørsmål

Om

Algoritmer, tillit og sandkasser

av Eirik Gulbrandsen | jan 6, 2021

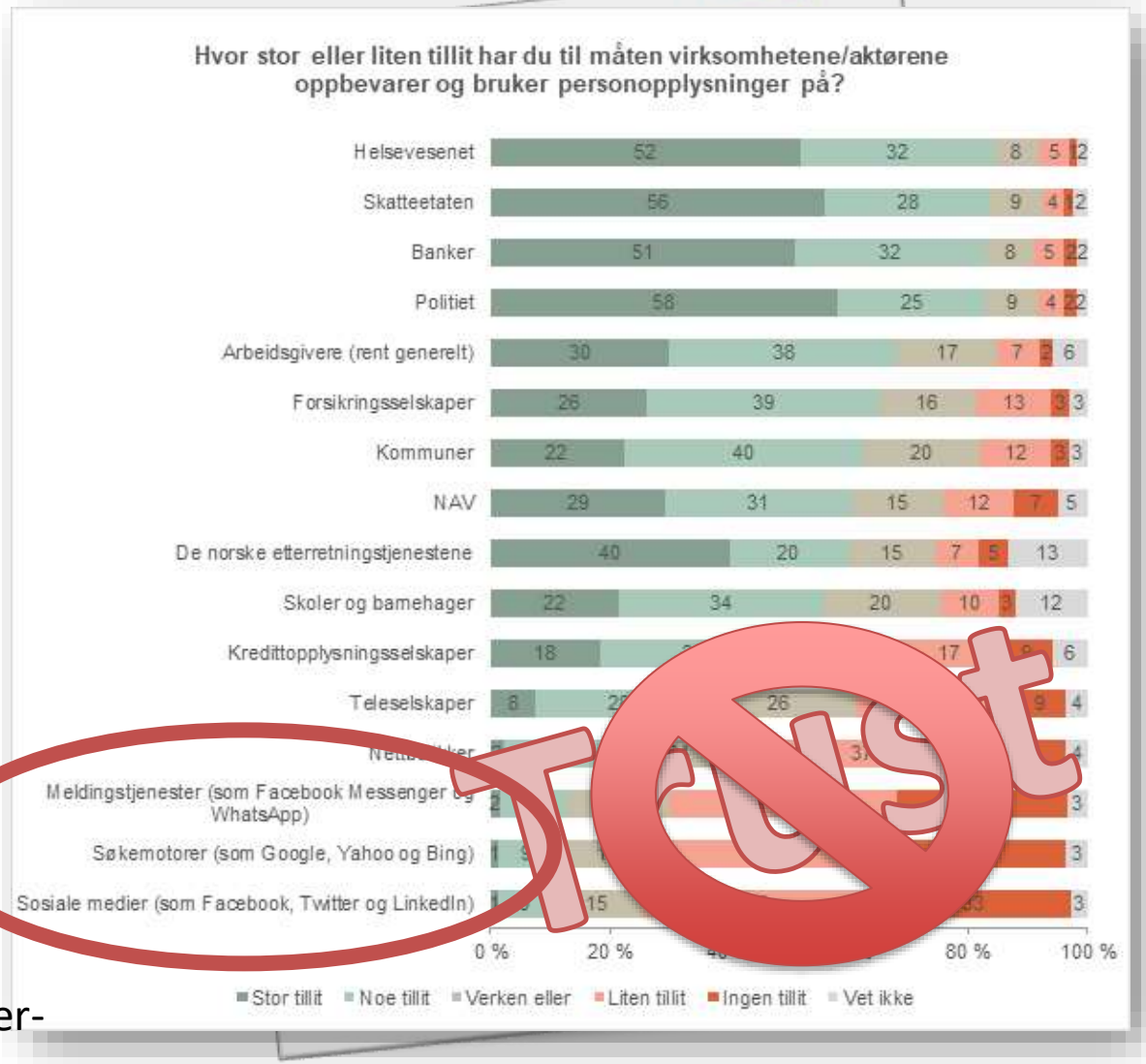
Mot slutten av 1800-tallet i USA undersøkte kjemikeren Harvey Wiley innholdet i en rekke industrielt produserte matvarer. 90 prosent av honningkrukkene inneholdt ikke honning. Lønnesirup var stort sett ikke lønnesirup. Og syltetøy bestod, som de nevnte matvarene, også stort sett av billig maissirup, 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 nedkjølingseffekt. Det betyr at mange begrenser bruken av digitale tjenester eller lar være å bruke dem. Effekten av digitalisering bremses fordi man ikke har den nødvendige tilliten til tjenestene – med rette.

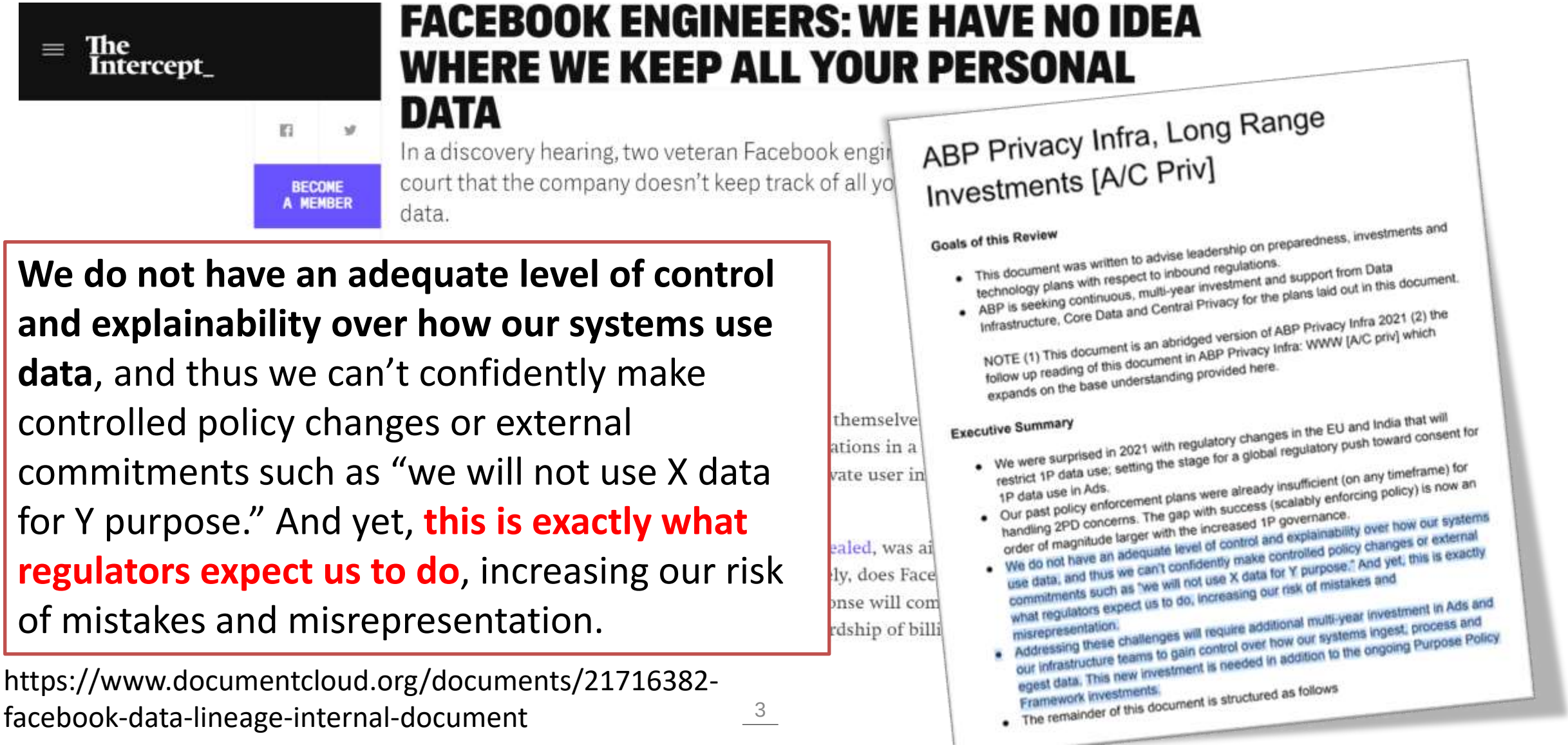
Det blir stadig tydeligere at gigantene bak «sosiale medier» – som 1800-tallets matprodusenter – ikke tilbyr tjenestene til samfunnets og enkeltmenneskenes beste. Google, Facebook mfl. er i realiteten først og fremst annonsesalgsselskaper. Gjennom kynisk bruk av algoritmer, ønsker de først og fremst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



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

Out of control → null kontroll…



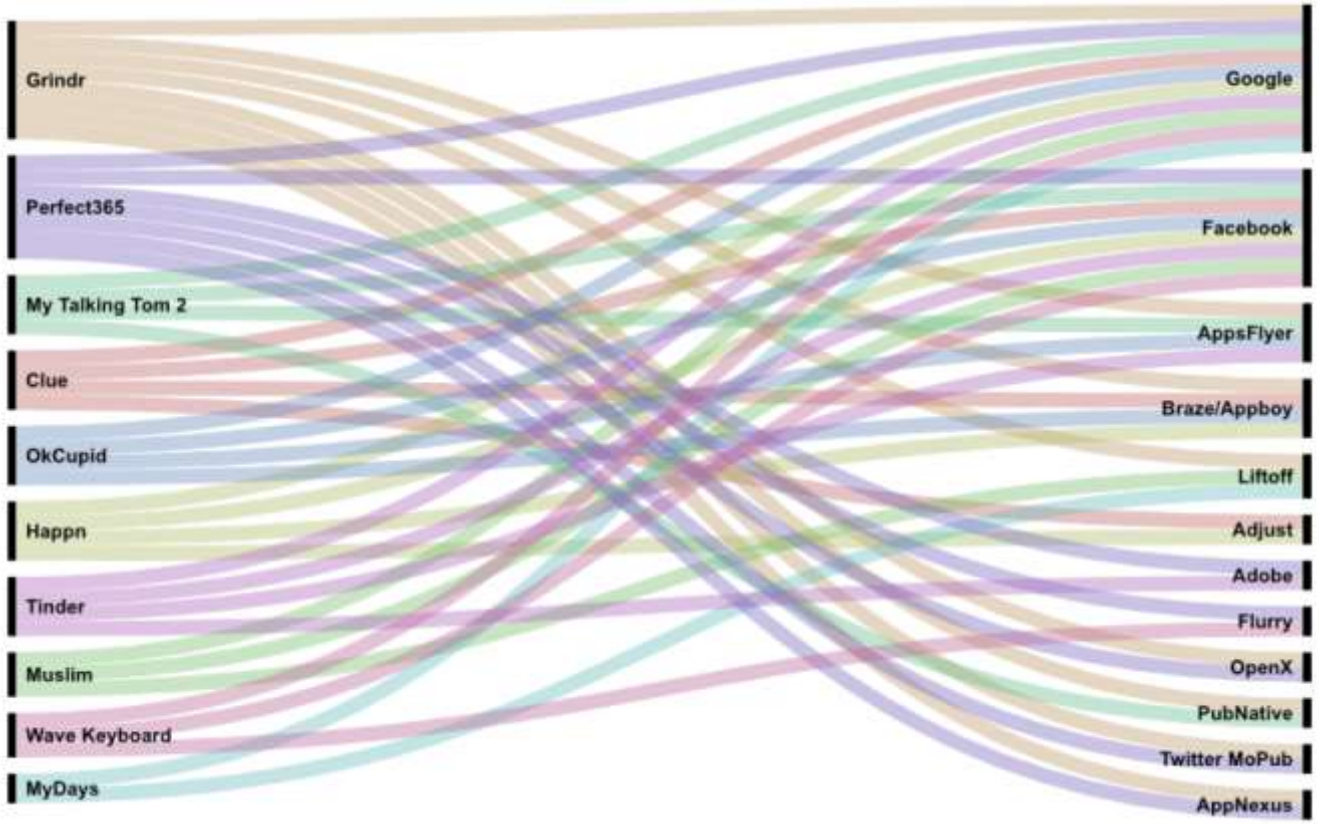
We do not have an adequate level of control and explainability over how our systems use data, and thus we can't confidently make controlled policy changes or external commitments such as "we will not use X data for Y purpose." And yet, **this is exactly what regulators expect us to do**, increasing our risk of mistakes and misrepresentation.

https://www.documentcloud.org/documents/21716382-facebook-data-lineage-internal-document

SocMed = ad market = personal data = Adtech

- «Adtech» carries SocMec
  - Grindr (plattforms; Google, Apple)
  - Ad networks (**Facebook**, Twitter/Mopub, Appnexus, etc)

Sends a lot of user data, including GPS location, IP address, gender, and age, to a large variety of third parties including **AdColony, AppNexus (Xandr)**, **Bucksense**, **MoPub, OpenX**, **PubNative**, and **Smaato**. Uses **MoPub** for ad mediation. Sends user information including "relationship type" to **Braze**.



TECHNICAL REPORT
"OUT OF CONTROL" – A REVIEW OF DATA SHARING BY POPULAR MOBILE APPS

SocMed = ad market = personal data = Adtech

- «Adtech» carries SocMec
  - Grindr (plattforms; Google, Apple)
  - Ad networks (**Facebook**, Twitter/Mopub

Sends a lot of user data, including GPS location, IP address, gender, and age, to a large variety of third parties including **AdColony, AppNexus (Xandr)**, **Bucksense**, **MoPub, OpenX**, **PubNative**, and **Smaato**. Uses **MoPub** for ad mediation. Sends user information including "relationship type" to **Braze**.

TEK.NO TESTER ARTIKLER TOPPLISTER TJENESTER

Datatilsynet vil ha slutt på Metas målrettede reklame

Meta kan få en million kroner i dagbøter om de ikke retter seg.

digi.no Tekjobb Nyhetsbrev Tips oss

– Vi føler oss litt som David mot Goliat

Det irske datatilsynet mener Meta bryter loven, men vil ikke håndheve reglene. Derfor har Datatilsynet tatt saken i egne hender.

SocMed = ad market = personal data = Adtech

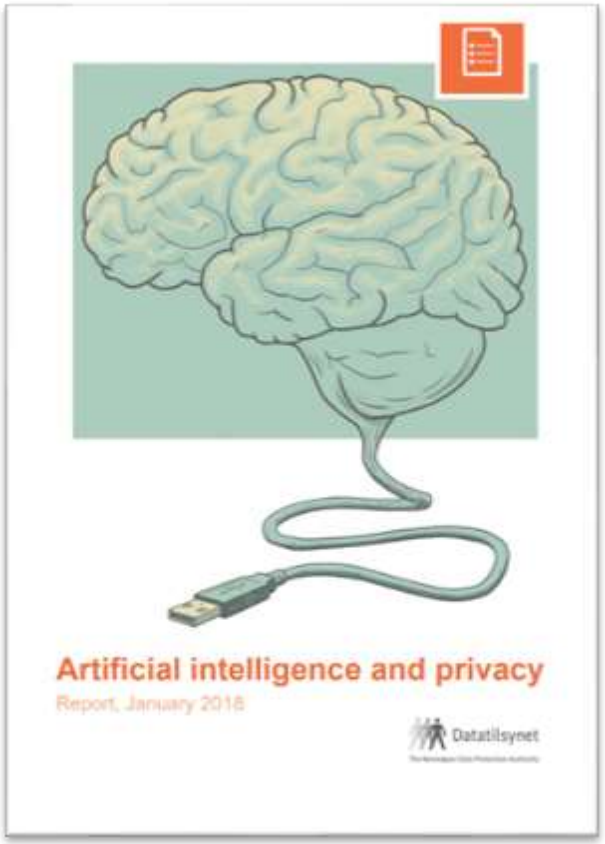
DSA - mer 45 millioner månedlige brukere i EU:

Alibaba AliExpress, Amazon Store, Apple AppStore, Booking.com, Facebook/meta, Google Play, Google Maps, Google Shopping, Instagram, LinkedIn, Pinterest, Snapchat, TikTok, Twitter/X, Wikipedia, YouTube, Zalando, Bing, Google Search

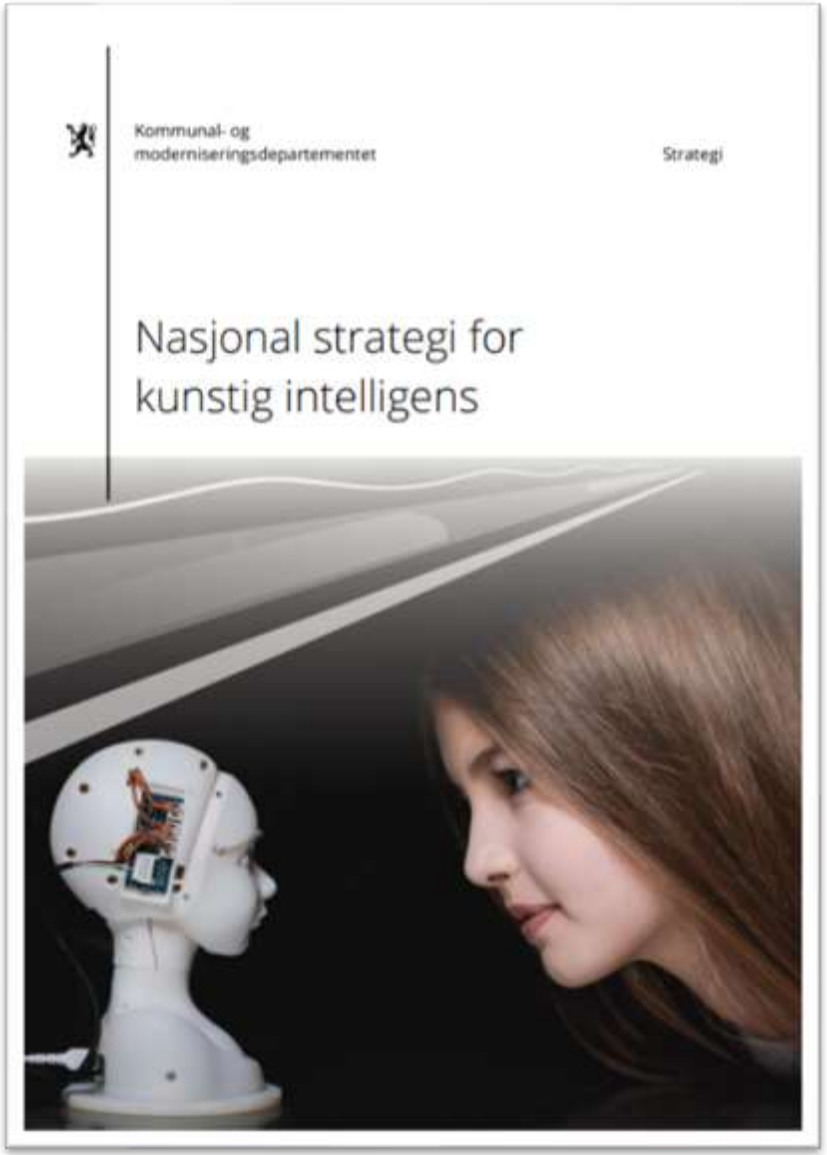
- ❑ Illegal content, transparent advertising and disinformation

Datatilsynets arbeid med kunstig intelligens

- KI og konsenser for personvern har vært et viktig tema i flere år
- Nøkkelutfordringer: hvordan regulere og trygge retten til personvern og sikre at KI er etisk og ansvarlig



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.»

«…etablere en regulatorisk sandkasse for personvern under Datatilsynets myndighetsområde.»

Hva ønsker vi å oppnå i sandkassen?

Virksomheter

- Økt **forståelse for de regulatoriske kravene.** Løsninger utviklet i sandkassen vil kunne fungere som **foregangseksempler**
- Gir mulighet til å avdekke eventuelle **svakheter og sårbarheter på et tidlig stadium** i prosessen – innebygd personvern i praksis.

Datatilsynet

- Øke tilsynets **kunnskap og forståelse av KI-drevne løsninger.**
- Utarbeide **veiledningsmateriale** basert på erfaring med utvikling av løsninger i sandkassen.

Enkeltindivider og samfunnet

- **Bygge tillit** til nye KI-løsninger ved at **utvikling av nye og innovative løsninger** foregår innenfor **ansvarlige rammer**.
- Forbrukere vil kunne dra **nytte av nye tjenester og produkter**, samtidig som viktige **personvernhensyn er ivaretatt.**

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: 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

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

Forklarbarhet / XAI av lærende systemer og GDPR

Artikkel 15

Den registrertes rett til innsyn

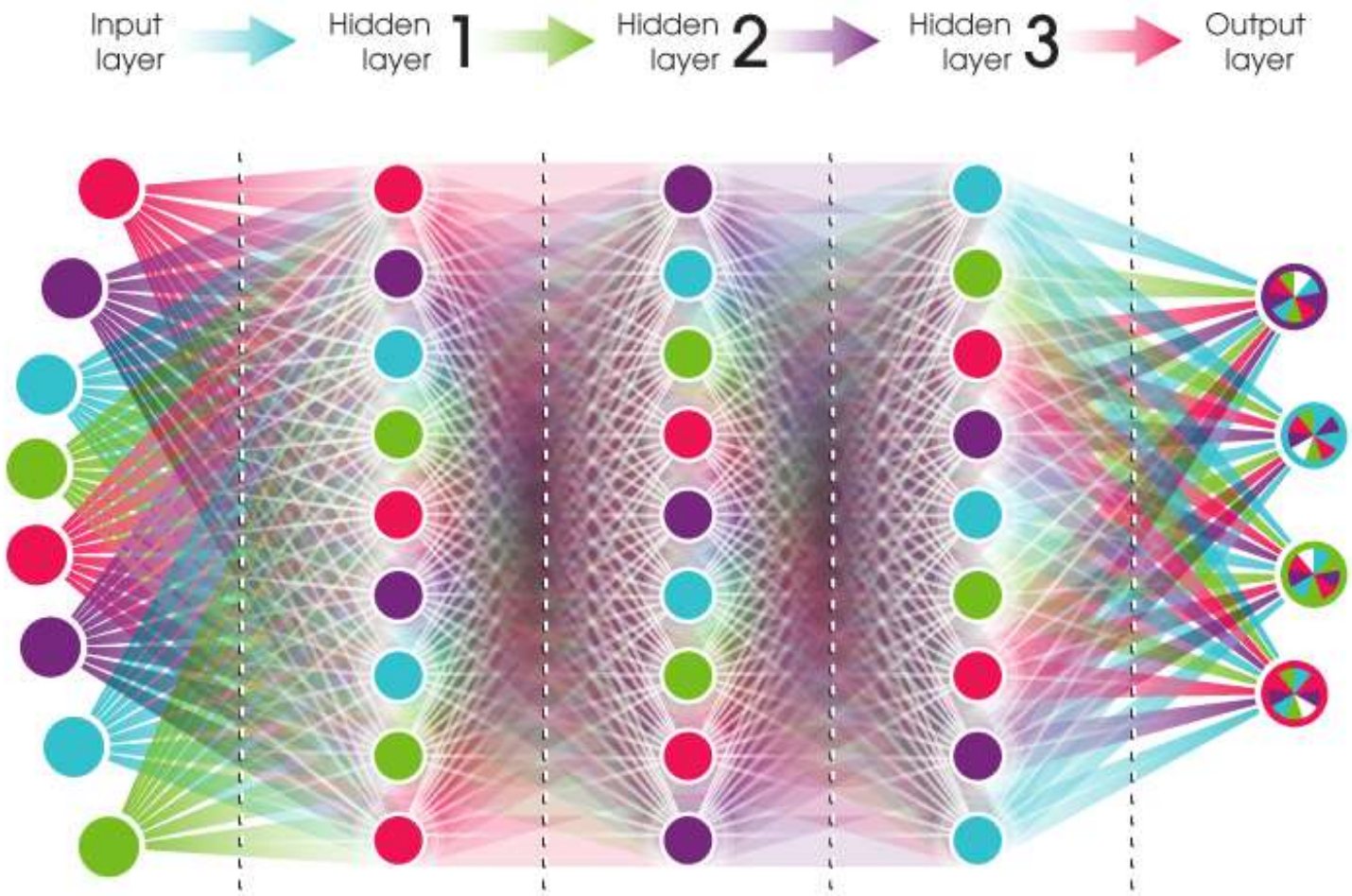
…relevant informasjon om den underliggende logikken samt om betydningen og de forventede konsekvensene…

Artikkel 25

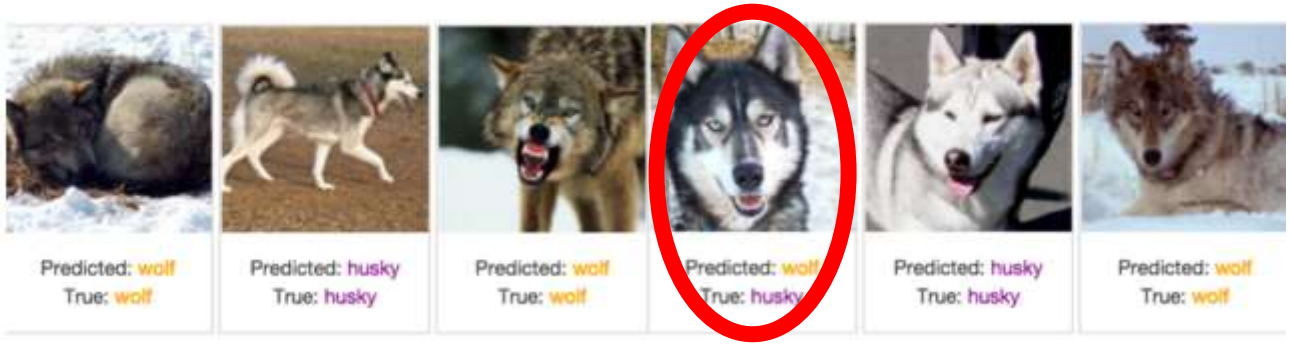
Innebygd personvern og personvern som standardinnstilling

…integrere de nødvendige garantier i behandlingen for å oppfylle kravene i denne forordning…

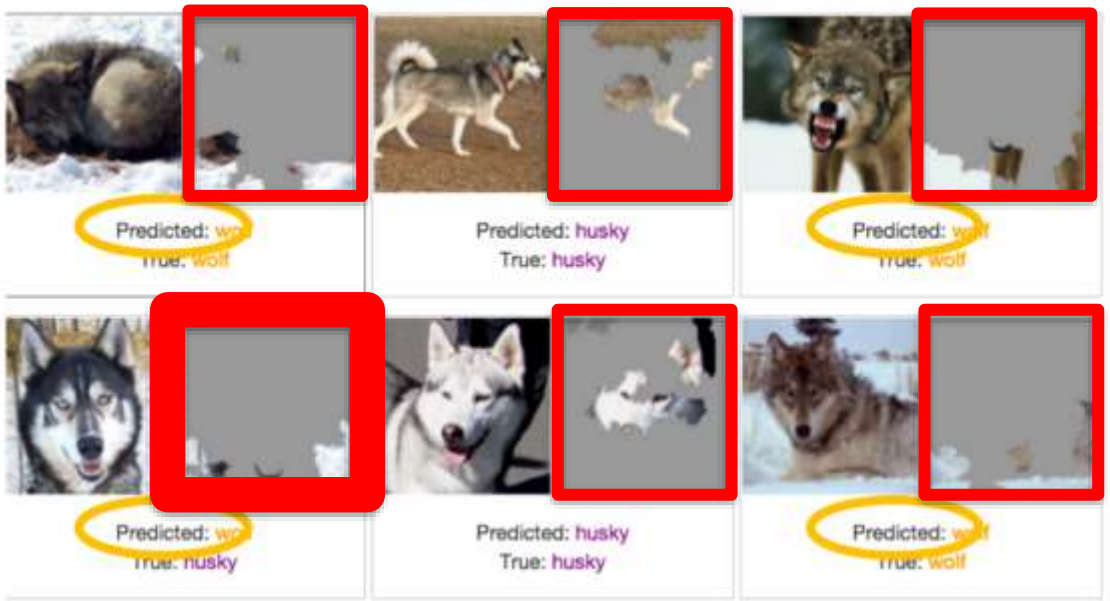
Forklarbarhet / XAI av lærende systemer



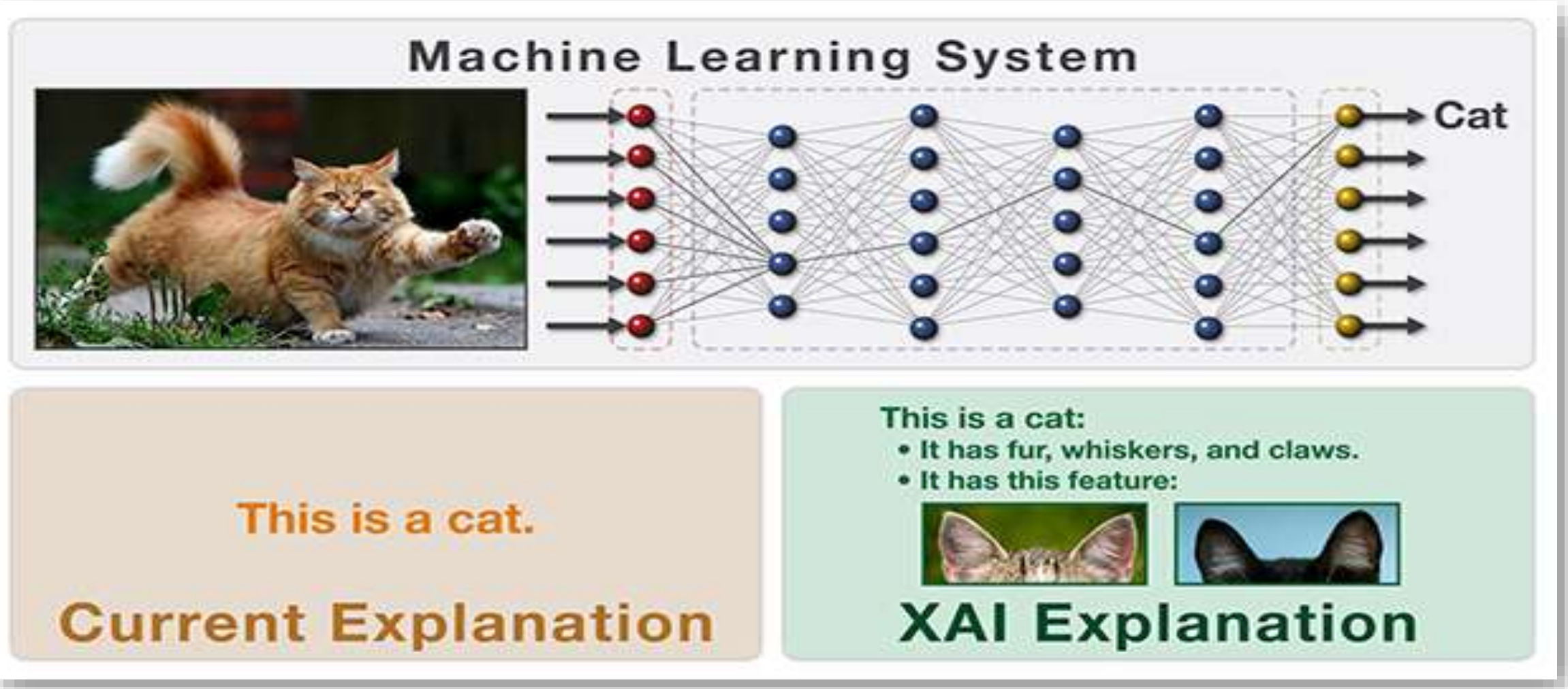
wolf v.s. husky



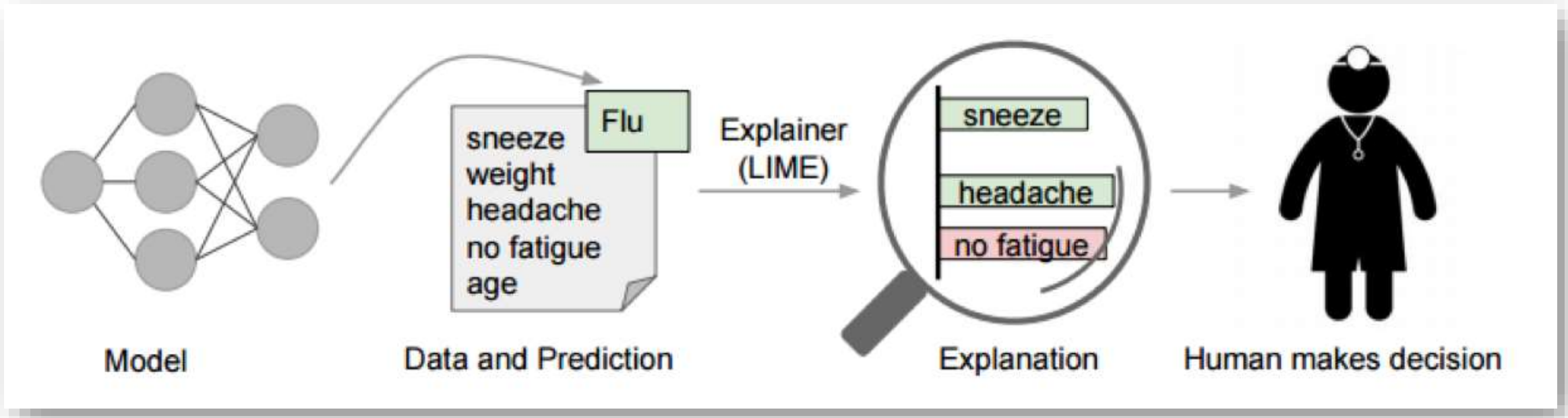
En snøpredikator…



Forklarbarhet / XAI av lærende systemer



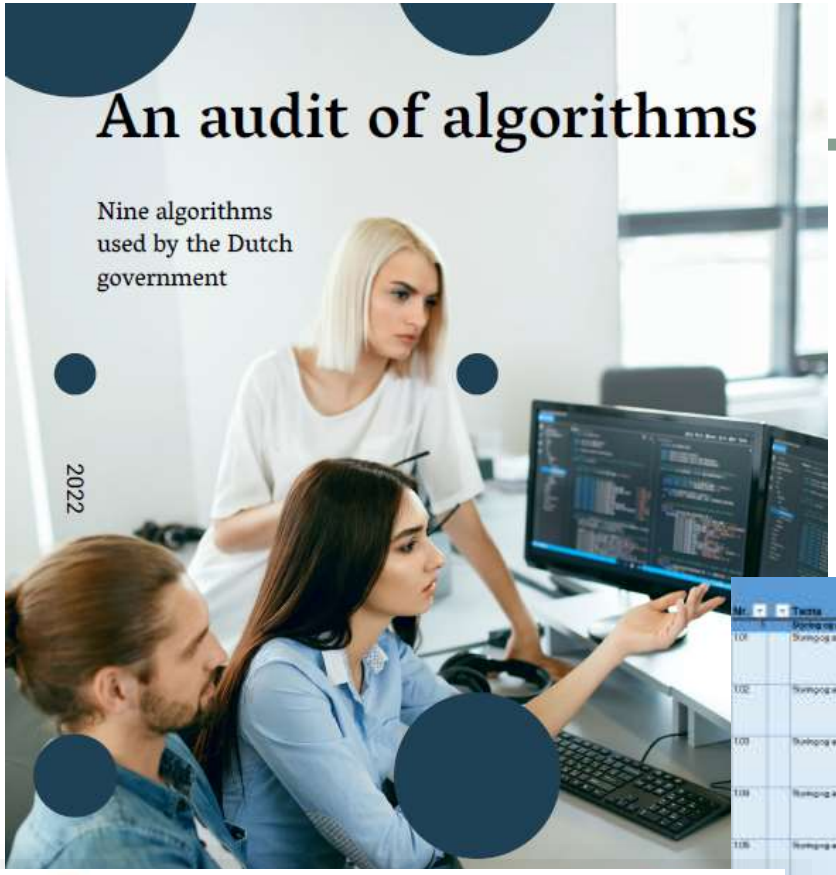
Forklarbarhet / XAI av lærende systemer



LIME – Local Interpretable Model-Agnostic Explanations

Algoritmetilsyn

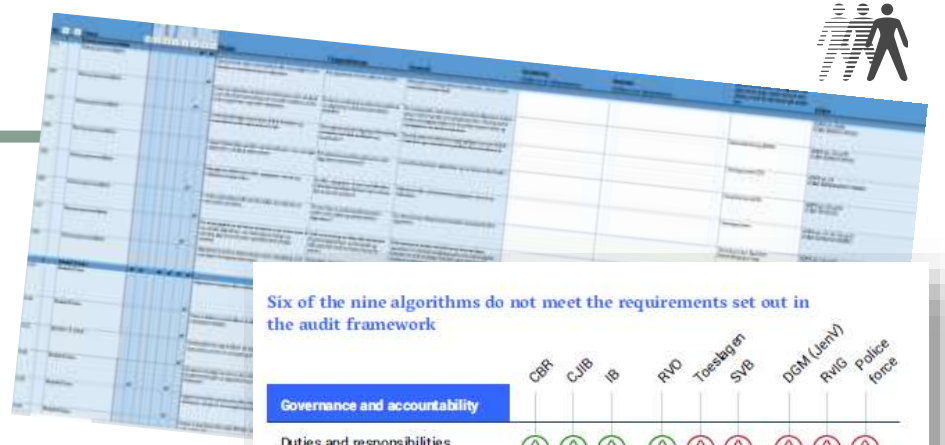
…av (maskin)lærende systemer som benytter personopplysninger.



ico.

Information Commissioner's Office

Algoritmetilsyn – eksempler

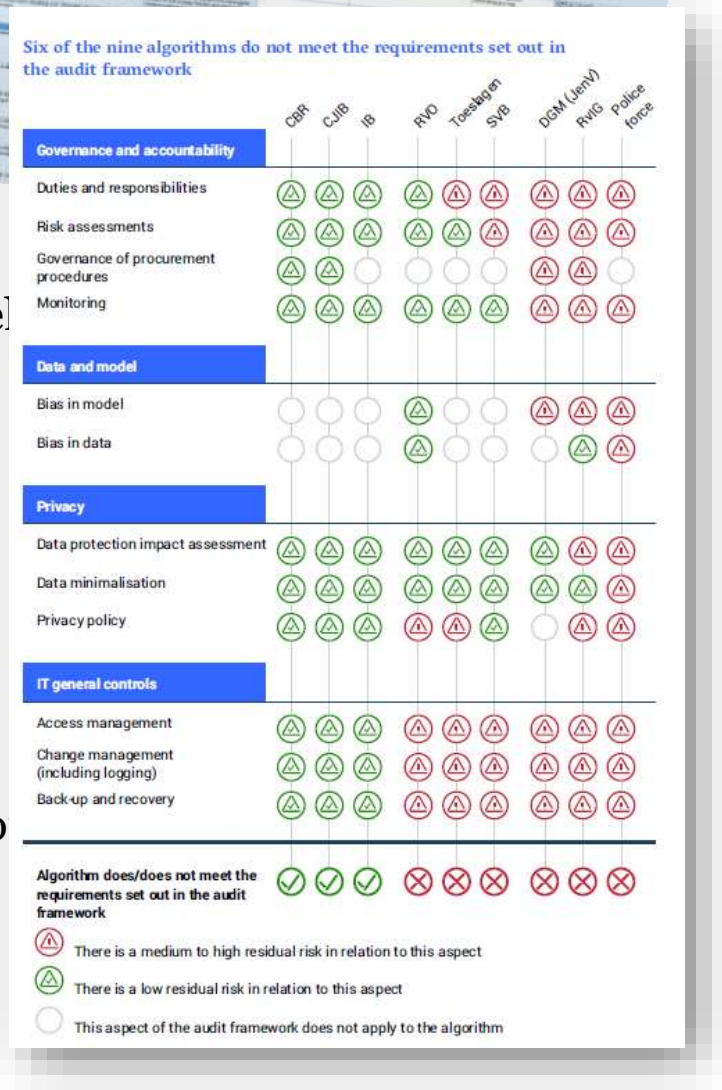


1. **Styring og ansvarlighet**
   - ✓ Har algoritmen et klart definert formål?
   - ✓ Er roller, oppgaver og ansvar (inkludert eierskap) definert?

2. **Modell og data (XAI)**
   - ✓ Er algoritmen forklarbar og er det forsøkt å finne en balanse mellom model
   - ✓ Har opplærings-, test- og valideringsdata blitt behandlet separat?

3. **Personvern (GDPR)**
   - ✓ Er det utført en personvernkonsekvensvurdering?
   - ✓ Er virkningen av bruken av algoritmen tydelig for registrerte?

4. **Informasjonssikkerhetsstyring**
   - ✓ Sjekkes det om tilgangsrettigheter er oppdaterte med tanke på miljøet algo
   - ✓ Er endringer gjort i koden/hyperparametre til algoritmen sporbare?



«AI Act» - Personvernperspektivet

I forhandlinger; Europakommisjonen, Europarådet og Europaparlamentet

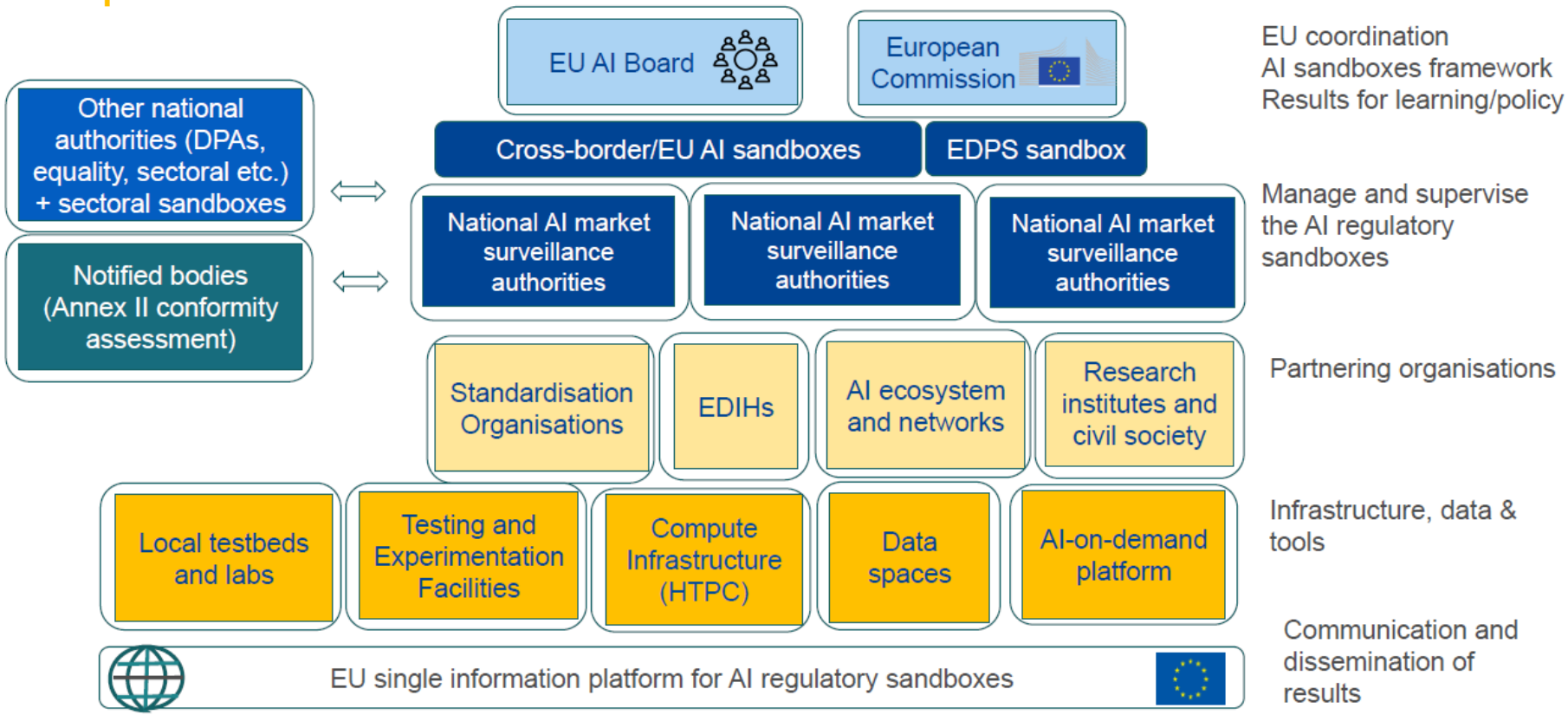
AI Act og sandkasser…

MEASURES IN SUPPORT OF INNOVATION

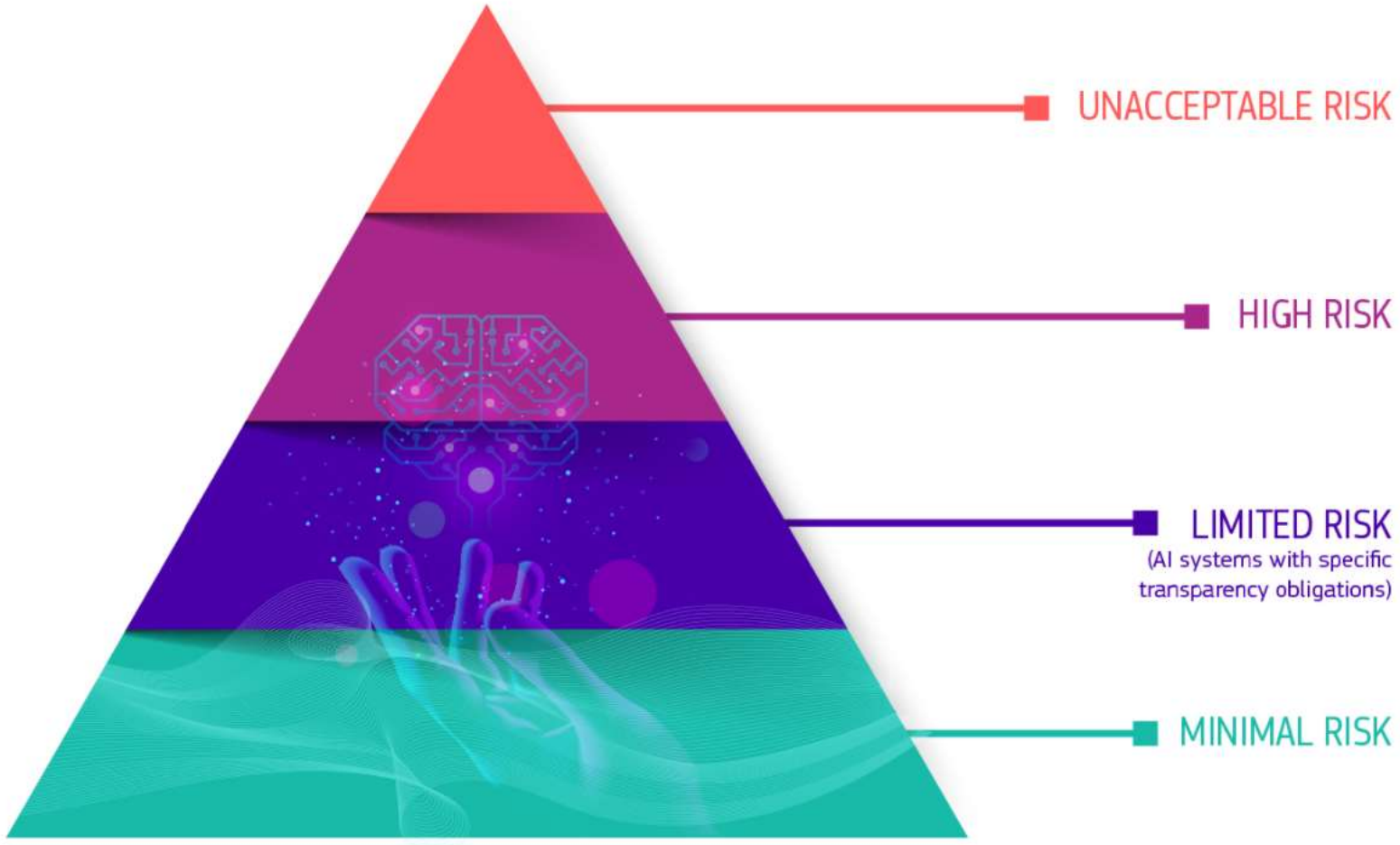
Article 53

AI regulatory sandboxes

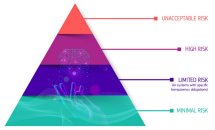
Governance of AI sandboxes



Fire kategorier KI



Høyrisiko-KI (og sandkasseprosjekter)

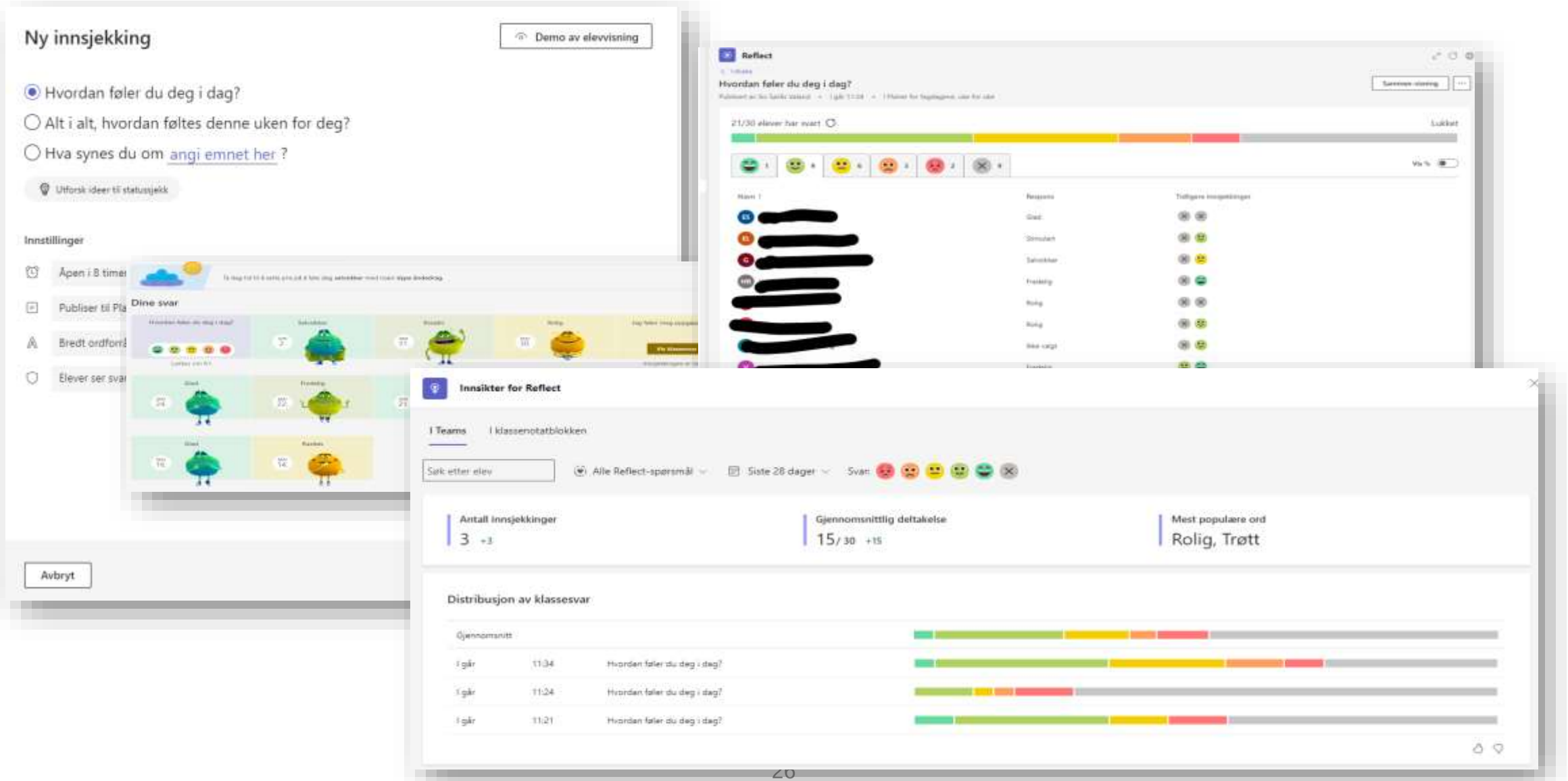


- Biometrisk identifikasjon (Mobai)
- Som sikkerhetskomponent i styring av trafikk, vann, varme og elektrisitet
- Opptak til utdanning og karaktersetting (AVT)
- Rekruttering og arbeidsforhold (Secure Practice, Simplifai)
- Adgang til offentlige tjenester og ytelser (NAV, AHUS, Helse Bergen)
- Kredittvurdering (Finterai)
- Uttrykning av nødtjenester
- Justisfeltet (politi, immigrasjon og domstoler) (PHS)
- Som sikkerhetskomponent (for eksempel i robotkirurgi)
- Listen kan endres av Kommisjonen – skal oppdateres én gang i året

EDPBs syn

- Totalforbud mot 'social scoring'
- Totalforbud mot å bruke AI til å kategorisere personer i sårbare grupper ut fra biometri
- Forbud mot å bruke AI til å predikere fremtidig adferd i justissektoren
- Forbud mot å bruke AI til å avdekke følelser

Microsoft Teams «Reflect» - kartlegging av følelser…



Big Brother Business

Om sikkerhets-myndighetenes kjøp og bruk av dine kommersielle persondata

Utgangspunkt

The New York Times

Intelligence Analysts Use U.S. Smartphone Location Data Without Warrants, Memo Says

The disclosure comes amid growing leg
the government uses commercially ava

Share full article

Gain unlimited access to all of The Times.

THE DAILY 202

A warr
buy yo

Analysis by C
with researcl

June 15, 2023 at 12:02 p.r

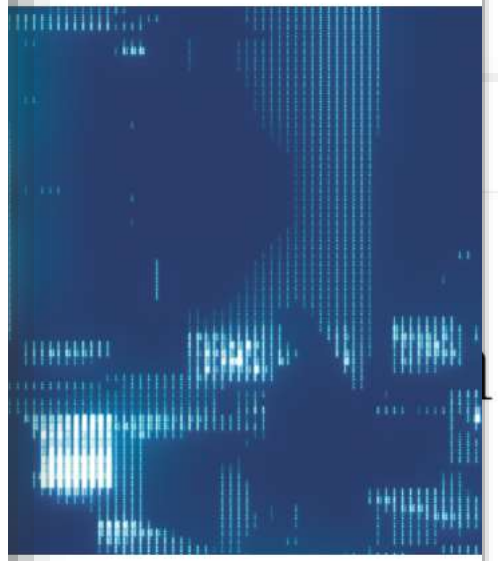
4.3 Om kjøp av metadata i bulk

Utvalget har stilt spørsmål til E-tjenesten om tjenestens hjemmelsgrunnlag for kjøp av metadata fra kommersielle aktører. E-tjenesten mente at enkelte anskaffelser av data fra kommersielle tilbydere ikke innebærer bruk av en inngripende metode etter e-loven kapittel 6.

Tjenesten anså derfor at forbudet mot innhenting i Norge etter e-loven § 4-1 ikke kom til anvendelse for den type anskaffelse av metadata som saken gjaldt. E-tjenesten mente at det er måten tjenesten kommer i besittelse av dataene på som er avgjørende for om anskaffelsen faller inn under e-loven kapittel 6 eller ikke.

Det sentrale spørsmålet for utvalget var om kjøp av metadata i bulk som inneholder personopplysninger utgjør et inngrep i enkeltpersoners privatliv. Rettspraksis fra Den europeiske menneskerettighetsdomstolen (EMD) viser at metoder som utgjør et slikt inngrep, krever klar forankring i lov.

I forarbeidene[7] til e-loven skrev Forsvarsdepartementet (FD)



RSMELDING 2022

DOKUMENT 7:1 (2022–2023)

Ny etterretningslov

DET KONGELIGE
FORSVARSDEPARTEMENT

Prop. 80 L

(2019–2020)

Proposisjon til Stortinget (forslag til lovvedtak)

Lov om Etteretningstjenesten

10.5.7.2 Passiv opptreden

Innhentingen forutsetter passiv opptreden, og skjer i ma

§ 6-2. *Åpne kilder*

Etterretningstjenesten kan innhente åpent tilgjengelig informasjon. Informasjon er ikke åpent tilgjengelig dersom tilgang krever aktiv fordekt opptreden eller forsering av passord eller lignende beskyttelsesmekanismer.

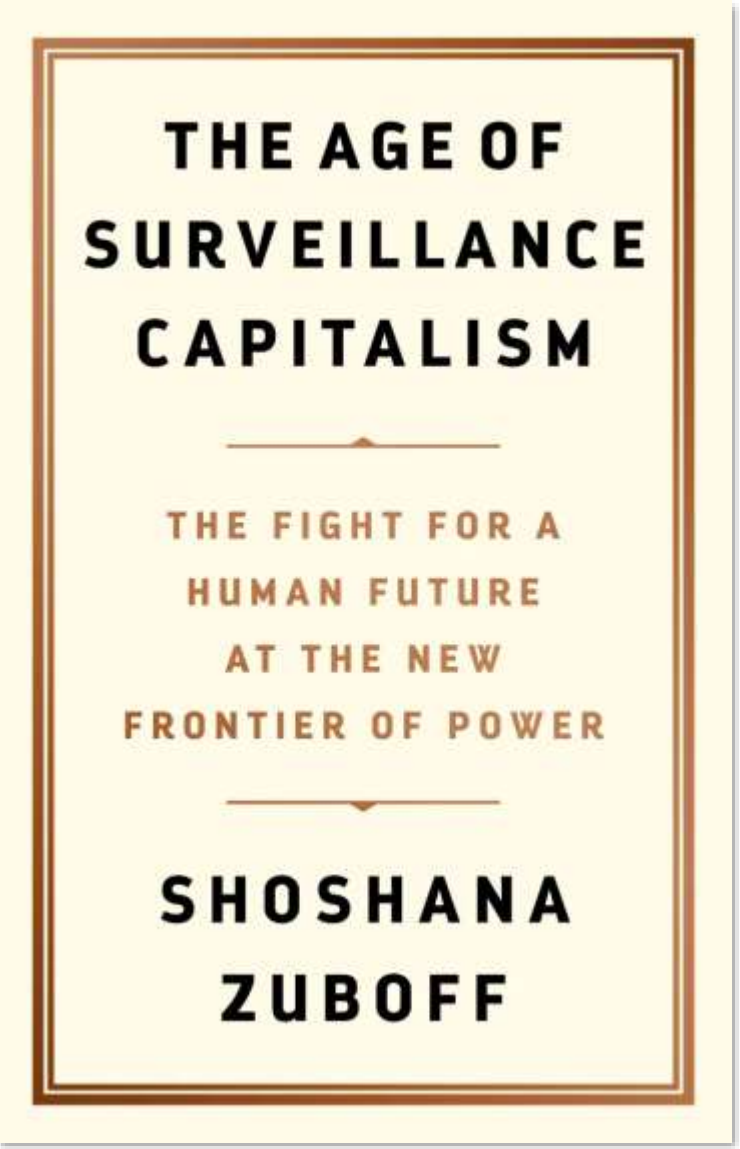
nødvendig å manipulere noen eller noe for å få den øns
være utenfor åpne kilder som innhentingsmetode. Det v

Side 171 av 395

Høring - Forslag til ny lov om Etterretningstjenesten
12.11.2018

Overvåkingskapitalismen

"State of exception" created by the Government use of special extra-legal prerogatives in the aftermath of 9/11 "favored Google's growth and the successful elaboration of its surveillance-based logic of accumulation"



Overvåkingskapitalismen

"State of exception" created by the Government use of special extra-legal prerogatives in the aftermath of 9/11 "favored Google's growth and the successful elaboration of its surveillance-based logic of accumulation"

Myndighetenes ønske om data sammenfaller ofte med private aktørers…

- IP-adresser
- Flypassasjerer
- DNA-databaser

Ikke et spørsmål og kapasitet, men om lovgivers valg

KI, samfunnssikkerhet og personvern

- **Personvern, sikkerhet, samfunnssikkerhet og tillit henger sammen for å hente ut gevinst fra KI og ivareta grunnleggende rettigheter**
- **Store og små aktører trenger (ønsker) både regulative krav og veiledning**
- **Dialogbasert veiledning er et godt verktøy (eks. sandkassemodellen)**
- **Algoritmetilsyn (også med lærende systemer) er et viktig styringsverktøy**
- **Også myndigheter* trenger tilsyn og ettergåelse, lovgivere trenger kunnskap og informasjon**

(inkludert norske…)*