Digitale spøkelser

KI-agenter uten kontroll i kommunale prosesser

Sara Rossebø | Rådgiver
Eirik Gulbrandsen | Spesialrådgiver
Seksjon Teknologi, Sikkerhet og Tilsyn

Datatilsynet

Problemstillingene i avviksmeldingene

- Datatilsynet har i økende grad mottatt avviksmeldinger knyttet til bruk av KI.

- Problemstillingene i disse kan grovt sett kategoriseres i to typer:
  1. Uautoriserte KI-verktøy som gjør opptak av og transkriberer digitale møter.
  2. Uautoriserte KI-agenter i nettleserutvidelser som omgår tilgangsstyring.

- Vi skal se på noen konkrete eksempler på avviksmeldinger og hvilke tiltak som ble iverksatt.

1. Uautoriserte KI-verktøy som transkriberer og gjør opptak av digitale møter

KI-verktøy som transkriberer og gjør opptak

- Møtedeltaker som har samtykket til at en tredjeparts KI-applikasjon kan benyttes uten at de andre møtedeltakerne har blitt informert eller det foreligger et behandlingsgrunnlag.

- Det er først når møtet avsluttes og møtedeltakerne eventuelt mottar en automatisk generert epost at man blir klar over at møtet er blitt tatt opptak av og transkribert.

Avviksmelding 1

«Det er avdekket betydelige utfordringer knyttet til bruk av tredjepartsverktøy, som Read AI, i møter der konfidensiell informasjon og personopplysninger behandles uten databehandleravtale. Gjennomgang av tjenestens vilkår viser at møteinnhold (lyd, bilde, tekst) behandles og lagres i USA, deles med underleverandører og markedsføringspartnere, og innebærer atferdsanalyse av deltakere. Avviket oppsto som følge av et praktisk behov for oversettelsesløsninger i møter [...]. For å imøtekomme dette behovet ble Read.AI tatt i bruk uten anskaffelsesprosess eller vurdering fra IKT, Personvern og Informasjonssikkerhetskoordinator, osv.».

Avviksmelding 2

«To av våre ansatte brukte selv Read.ai i flere møter uten å tenke seg om, herav et internt møte og to møter med eksterne der vedkommende var deltaker (ikke arrangør). Deltaker fra det første eksterne møtet, har i etterkant mottatt den auto-genererte mailen [og] innsett at dette ikke var i tråd med GDPR [...]. I transkriberingen vil alle personer som snakket i møte, bli gjengitt med fullt navn og sitert ordrett (i den grad teknologien klarer å sitere korrekt) i rapporten. Det kan derfor være sensitive opplysninger her.»

KI-verktøy som transkriberer og gjør opptak

- Mange av avviksmeldingene nevner Read.AI som verktøy, men det finnes flere.

- Denne sender ut varsel til alle møtedeltagerne om at opptak og transkripsjon er blitt gjort og at hver enkelt kan se disse om de logger seg inn på plattformen.

- Alle som da klikker seg inn blir automatisk innrullet som brukere av Read AI og drar tjenesten med seg videre i fremtidige Teams-møter.

KI-verktøy som transkriberer og gjør opptak

- Problematikken belyser er et større problemkompleks knyttet til tilgangsstyring og kontroll med applikasjoner, hvor ansatte ***kan logge på, integrere og gi tilgang til*** tredjepartstjenester uten at dette styres eller godkjennes.

Hva kan være konsekvensen?

ILLUSTRASJON: Omtrent 17.000 skal være rammet. Foto: Kristoffer Søvik

Stort personvernbrudd ved UiO – flere tusen rammet

Avviket har utviklet seg over tid og fått et stort omfang, skriver UiO.

https://www.universitas.no/datatilsynet-personvern-personvernbrudd/stort-personvernbrudd-ved-uio-flere-tusen-rammet/392029

Ikke-godkjente tilganger

Men hva gikk egentlig galt? Slik har UiO beskrevet det selv:

Studenter og ansatte kan knytte eksterne applikasjoner til Microsoft365-brukerne deres. Omtrent 1400 slike tjenester har hatt tilgang til brukernes data. Flere er godkjente.

Problemet: Et ukjent antall tjenester var ikke godkjent og kunne dermed ha uautorisert tilgang til personopplysninger.

En av disse er ReadAI: En slags digital møteassistent. Tjenesten kan blant annet transkribere og oppsummere møter i Google Meet, står det på ReadAIs nettsider.

UiO opplyste i juni Datatilsynet om at 59 brukere hadde brukt tjenesten. Dette var starten på undersøkelsene som avdekket avvikets store omfang.

– Ansatte og studenter har selv hatt mulighet til å samtykke til at tredjepartstjenester har fått tilganger inn i data som er lagret i UiOs Microsoft 365, skriver assisterende universitetsdirektør Johannes Falk Paulsen i en e-post til Universitas.

UiO har manglet rutiner for å godkjenne at disse har blitt knyttet til Microsoft365.

Svikten skal ha pågått siden 2014.

Usikkert omfang

Omtrent 17.000 UiO-brukere skal ha samtykket til at ikke-godkjente tredjepartstjenester har fått tilgang til data.

2. Uautoriserte KI-agenter i nettleserutvidelser som omgår tilgangsstyring

Innebygde KI-agenter som omgår tilgangsstyring

- KI integreres i økende grad i nettlesere i form av innebygde KI-agenter: eksempelvis Copilot i Edge.

- Brukere kan be Copilot om å basere sine svar på innholdet som den har tilgjengelig i nettleseren.

- Virksomheter har i økende grad inngått Enterprise-avtaler med Microsoft slik at man kan ta i bruk denne tjenesten på en måte som sikrer bedre kontroll på behandling, sletting og lagring av data.

- Men vi har i økende grad fått avviksmeldinger om ansatte som har tatt i bruk den innebygde gratisversjonen av Copilot på områder hvor dette ikke er risikovurdert eller godkjent.

Avviksmelding 3

«Vårt dokumentsenter har oppdaget at det er mulig for Copilot integrert i Edge å analysere/oppsummere tekst fra dokumenter i saks- og arkivsystemet, da denne er skybasert og åpnes i Edge. Copilot kan på denne måten ha fått tilgang til personopplysninger og/eller særlige personopplysninger gjennom denne integrerte funksjonen i Edge. […] Hendelsen hadde sitt utspring i en tilsiktet handling fra en ansatt, som i strid med opplæring og interne retningslinjer for bruk av Copilot M365, aktivt ba Copilot M365 om å oppsummere et dokument i saks- og arkivsystemet.»

Avviksmelding 4

«Enkelte ansatte har brukt det innebygde panelet for Copilot i Edge-nettleseren til å hjelpe til med saksbehandling i fagsystemet […]. De har åpnet dokumentet i [fagsystemet] og bedt Copilot sammenfatte dokumentet og/eller vurdere innholdet […]. Copilot ble brukt i tro om at dette er et integrert verktøy i fagsystemet.»

Innebygde KI-agenter som omgår tilgangsstyring

- I denne innebygde gratisversjonen av Copilot er man ikke innlogget som bruker, og har derfor ikke kontroll på hva som skjer videre med opplysningene som legges inn i tjenesten.

- Konsekvensen er at de ansatte potensielt deler (sensitive) personopplysninger med en tredjepartsleverandør uten at det er inngått en databehandleravtale.

- Årsaken til slike avvik handler som regel om manglende sikkerhetsinnstillinger.

Innebygde KI-agenter som omgår tilgangsstyring

- Utfordringer og svakheter i ***den digitale grunnmuren***, som dårlig tilgangsstyring og kontroll med applikasjoner, blir i slike hendelser synlig og kraftig forsterket av KI-verktøy.

Tekniske og organisatoriske tiltak for å forhindre lignende hendelser

Read AI og andre tredjeparts KI-applikasjoner

- Mulig å blokkere tjenester som Read.AI i Teams Admin Center og/eller Entra ID.
- Mulig å blokkere eposter som kommer fra relaterte domener.
- Oppdatere Teams-policy, slik at det kun er host eller co-host som kan slippe deltakere inn i møter.
- **På et generelt nivå:** sperre for at ansatte kan laste ned tredjepartsapplikasjoner uten sentral godkjenning.

Copilot og KI-agenter i nettleserutvidelser

- Sette opp sikkerhetsregler som forhindrer at Copilot dukker opp i nettleseren.
- Følge opp endringer i funksjonalitet i programvarer som kommer med sikkerhetsoppdateringer fra Microsoft.
- Sikre gode opplæringsrutiner og bevisstgjørelse av problemstillingene for de ansatte. Det er virksomhetens ansvar at de ansatte har denne kunnskapen.
- Lage interne retningslinjer for bruk av KI.

Tiltak for å forhindre brudd på personopplysningssikkerheten

- Viktigheten av å sikre ***kontroll*** med tredjepartsapplikasjoner, nettleserutvidelser og hva sluttbrukere kan laste ned, aktivere og bruke av programvare.

- Tilgangskontroll er et nøkkelord her: still spørsmål til
  - Hvem som kan gjøre hva?
  - Hvem som har tilgang til hva?
  - Hvem som skal godkjenne hva?

- Dersom KI skal tas i bruk, gjøre risikovurderinger, inngå en databehandleravtale og vurder hvordan personvernet tas hensyn til i denne.

Datatilsynet, sandkasser, KI og språkmodeller

Er det så farlig da?

CURSOR

code llama

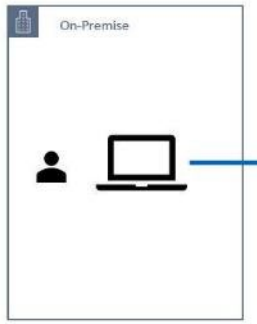
MISTRAL AI_

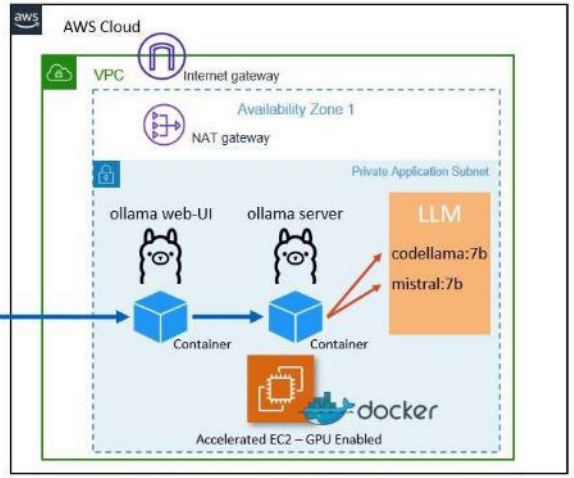
Gemini

Get up and running with large language models, locally.

Run Llama 2, Code Llama, and other models.
Customize and create your own.

Nor LLM





Er det så farlig da?

Lovlighet
Personopplysningssikkerhet
Ansvarlighet

Datatilsynets KI-sandkasse - siden 2020

NAV: Prediksjonsmodell for sykefravær

17.01.2022

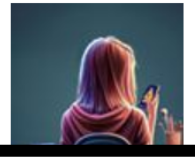
NAV h
til å f

AVT: Aktivitetsdata for vurdering og tilpassing

Politihøgskolen: PrevBOT

20.03.2024

Kunstig intelligens (KI) kan



Juridisk ABC: Jussboten LawAi

11.12.2024

Da Juridisk ABC lanserte LawAi i september 2023, var det med mål om å gjøre arbeidsrettslig veiledning enkel og tilgjengelig for alle gjennom generativ kunstig intelligens. Men hvordan kan man lage et generativt KI-verktøy for HR og arbeidsrett - som ofte involverer (sensitive) personopplysninger - uten å komme på kant med regelverket?

NTNU: Copilot med personvernbriller på

26.11.2024

Microsoft lanserte sin KI-løsning Copilot for Office-pakken i november 2023, med potensial for å forenkle arbeidshverdagen betraktelig. Noen har tatt det i bruk, helt eller delvis. Mange sitter på gjerdet. For hva skjer egentlig når du slår på Microsoft 365 Copilot?

personve
oppstart
kampen

intelligens i forbindelse med at de planlegge
bruke kunstig intelligens i sin app. I prosjekt
det diskutert hvordan Ruter kan være åpne
behandlingen av personopplysninger som vil skje i
løsningen, blant annet om formålene.

gjøre informerte valg når de skal kjøpe intelligente løsninger, som for eksempel en digital arkivmedarbeider (DAM).

Copilot-rapportene; Datatilsynet og NTNU

9. Små og kontrollerte steg

- Det er mulig å ta i bruk Copilot – MEN
  - Ikke for alle, og ikke for alt
- Vær aktsom – små steg
- Start med en utvalgt rolle med begrenset tilgang som utfører egnede behandlinger
- Lag et system for strukturert etterkontroll

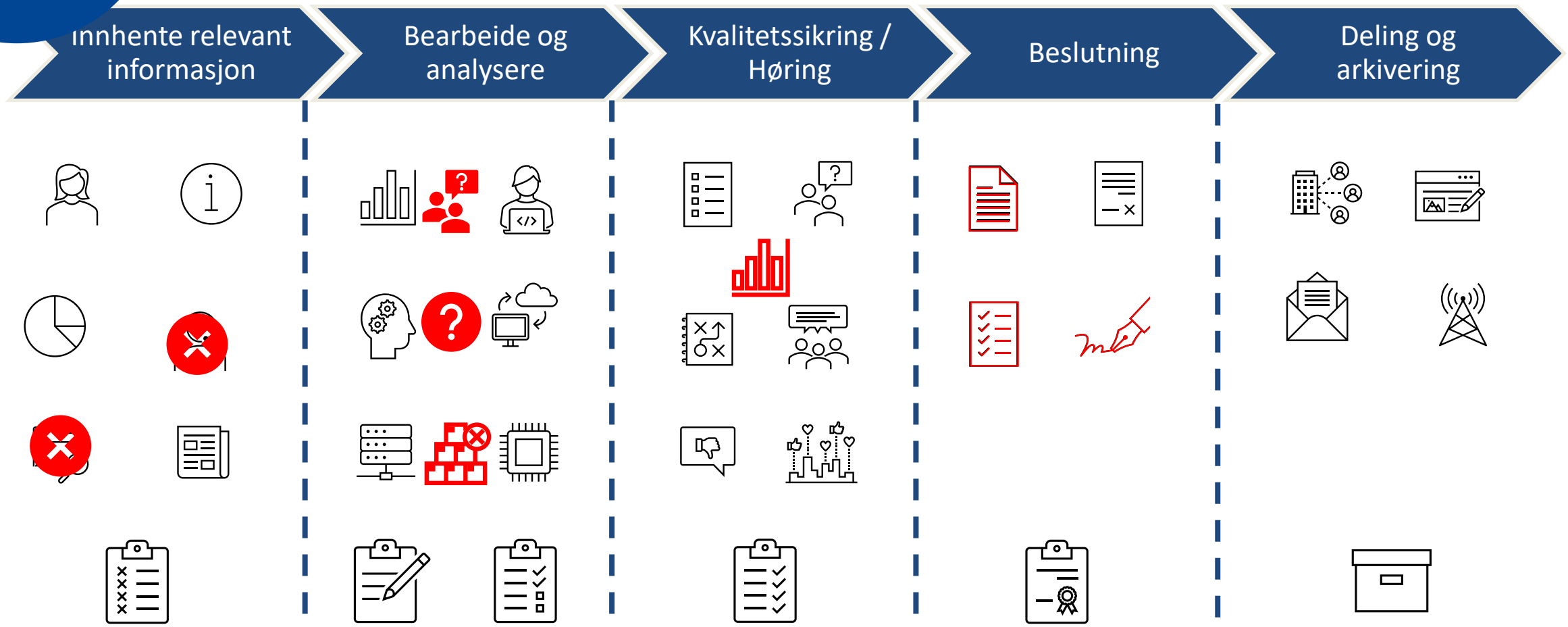
Hva har vi (NTNU) funnet ut?

1. *«Copilot er helt glimrende når du allerede kan det du vil at den skal hjelpe deg med»*
2. *«Copilot kan påvirke utøvelse av offentlig myndighet»*
3. *«Copilot behandler enorme mengder personopplysninger på nye og ukontrollerte måter»*
4. *«Microsoft 365 er utfordrende å forvalte»*
5. *«Copilot er tidlig i utviklingsløpet»*
6. *«Copilot påvirker organisasjonen»*
7. *«Copilot kan brukes til å overvåke og måle prestasjoner og adferd»*
8. *«Copilot fungerer tidvis skikkelig bra»*

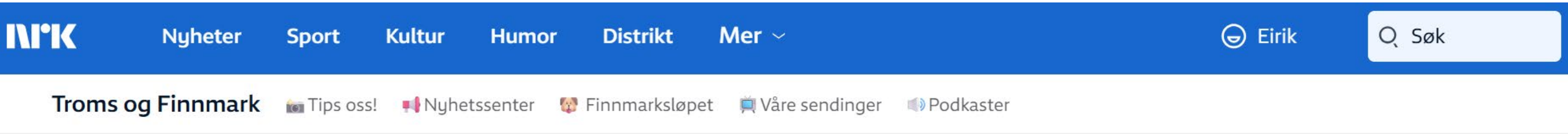
Hva har vi (NTNU) funnet ut?

1. *«Copilot er helt glimrende når du allerede kan det du vil at den skal hjelpe deg med»*
2. ***«Copilot kan påvirke utøvelse av offentlig myndighet»***
3. *«Copilot behandler enorme mengder personopplysninger på nye og ukontrollerte måter»*
4. *«Microsoft 365 er utfordrende å forvalte»*
5. *«Copilot er tidlig i utviklingsløpet»*
6. *«Copilot påvirker organisasjonen»*
7. *«Copilot kan brukes til å overvåke og måle prestasjoner og adferd»*
8. *«Copilot fungerer tidvis skikkelig bra»*

2 Saksbehandlingskjede



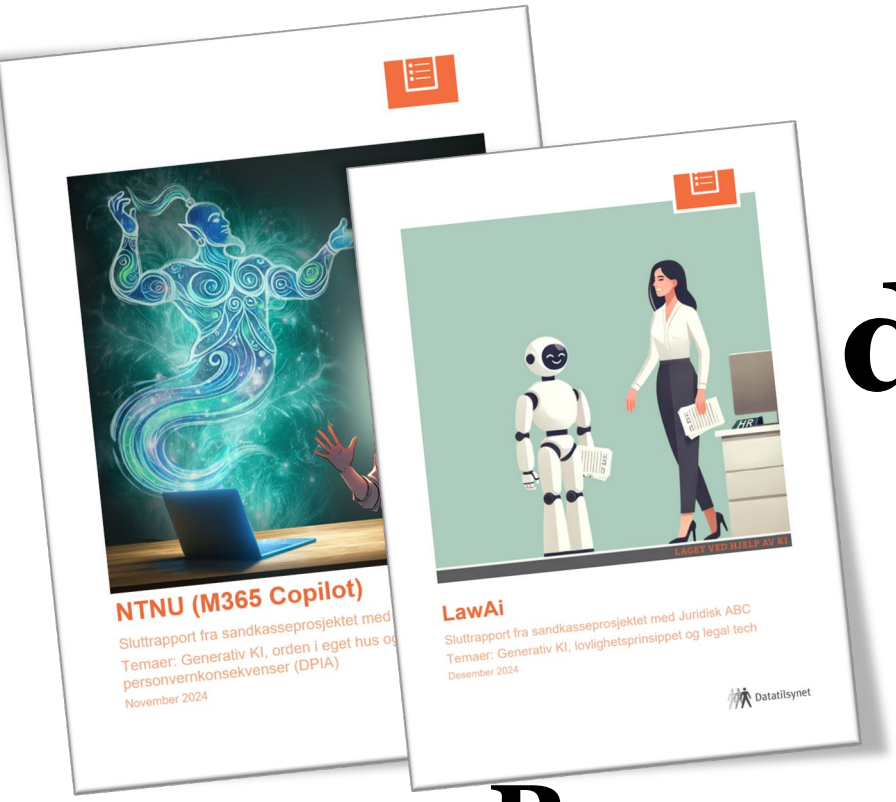
Google: «tromsø ki»



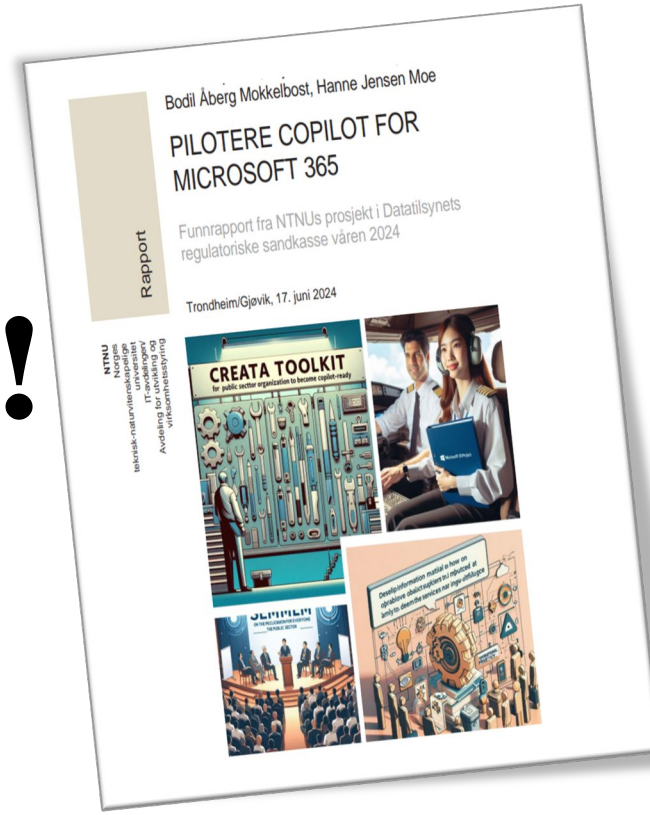
Kommune tatt for KI-bruk: – Dette er pinlig

Tromsø kommune brukte kunstig intelligens som hjelpemiddel i arbeidet med en viktig rapport. Rapporten inneholdt flere feil, noe KI-ekspert mener kunne vært unngått.

Oppvekstsjefen må gå av etter KI-skandalen i Tromsø



det er så farlig!



Lovlighet

Personopplysningssikkerhet

Ansvarlighet

NTNU lanserer egen KI-chat

Nå
spr
Diss

Mod
GPT-
Kimi
Mistr 3
GPT- Think

Begrenser tilgang til Copilot

Samtidig meldte NTNU i en pressemelding fredag at de nå begrenser tilgangen til Copilot Chat for ansatte.

Bakgrunnen er at Microsoft nylig har utvidet funksjonaliteten i Copilot Chat for Outlook. Dette krever nye vurderinger knyttet til personvern og sikkerhet før funksjonaliteten ruller ut.

Fra 19. mars 2026 var ikke Copilot Chat lenger tilgjengelig for ansatte via Outlook, mobilappen eller nettsiden m365.cloud.microsoft/chat.

Kilde: NTNU

MISTRAL AI_

Hey, I'm Lumo. Ask me anything. It's confidential.

Ask anything...

Upload | Web search | Press Enter to ask

Whatever you ask me is: Learn more ×

Private

Unlike other assistants, I don't record our conversations.

Safeguarded

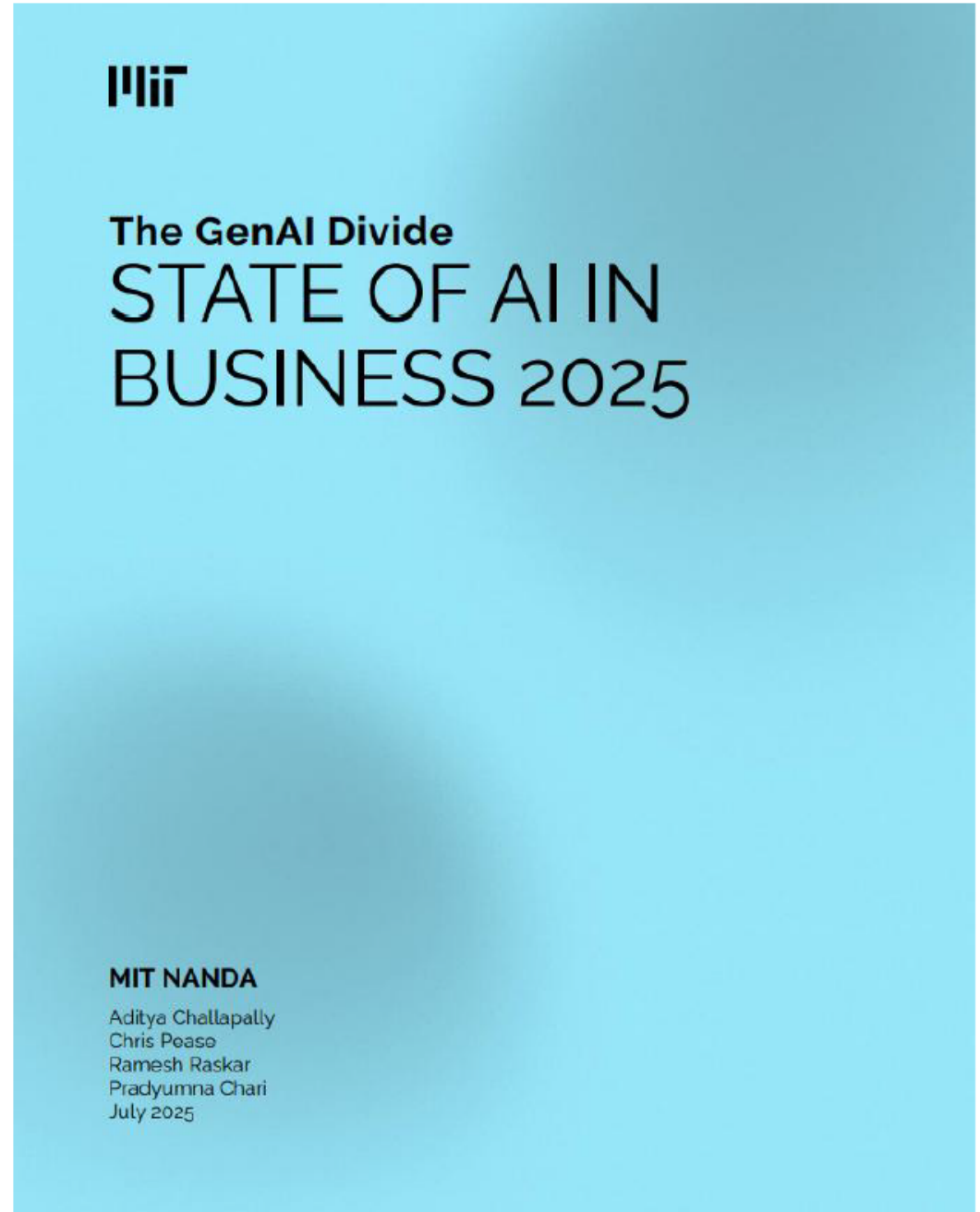
Not even Proton can access our chat history.

Treated with respect

Our conversations are never used for training.

Effektiviserende (en verdi)?

- 95% av virksomheter rapporterer INGEN gevinst
- (Til tross for investeringer for 30-40 milliarder dollar)
- Hva er reklame-pitchen og hva er reelt?
- Risiko vs. nytte vs. makt?



Eksempler på rammeverk for konkretisering av art 32-krav

Datatilsynets KI-sandkasse - siden 2020

NAV: Prediksjonsmodell for sykefravær

17.01.2022

AVT: Aktivitetsdata for vurdering og tilpassing

Politihøgskolen: PrevBOT

20.03.2024

Kunstig intelligens (KI) kan

Juridisk ABC: Jussboten LawAi

11.12.2024

Da Juridisk ABC lanserte LawAi i september 2023, var det med mål om å gjøre arbeidsrettslig veiledning enkel og tilgjengelig for alle gjennom generativ kunstig intelligens. Men hvordan kan man lage et generativt KI-verktøy for HR og arbeidsrett - som ofte involverer (sensitive) personopplysninger - uten å komme på kant med regelverket?

NTNU: Copilot med personvernbriller på

26.11.2024

Microsoft lanserte sin KI-løsning Copilot for Office-pakken i november 2023, med potensial for å forenkle arbeidshverdagen betraktelig. Noen har tatt det i bruk, helt eller delvis. Mange sitter på gjerdet. For hva skjer egentlig når du slår på Microsoft 365 Copilot?

bruke kunstig intelligens i sin app. I prosjekt

det diskutert hvordan Ruter kan være åpne

behandlingen av personopplysninger som vil

løsningen, blant annet om formålene.

gjøre informerte valg når de skal kjøpe intelligente løsninger, som for eksempel en digital arkivmedarbeider (DAM).

4. Evne til gjenoppretting av data og systemer

Har kommunen overordnede (styrende) retningslinjer for gjenoppretting av data og systemer?

Ja Nei

Hvis ja, inneholder retningslinjene:

a. Vurdering av systemenes kritikalitet?

Ja Nei

b. Prioriteringer av systemene?

Ja Nei

c. Krav til gjenopprettingstid?

Ja Nei

d. Fordeling av oppgaver og ansvar?

Ja Nei

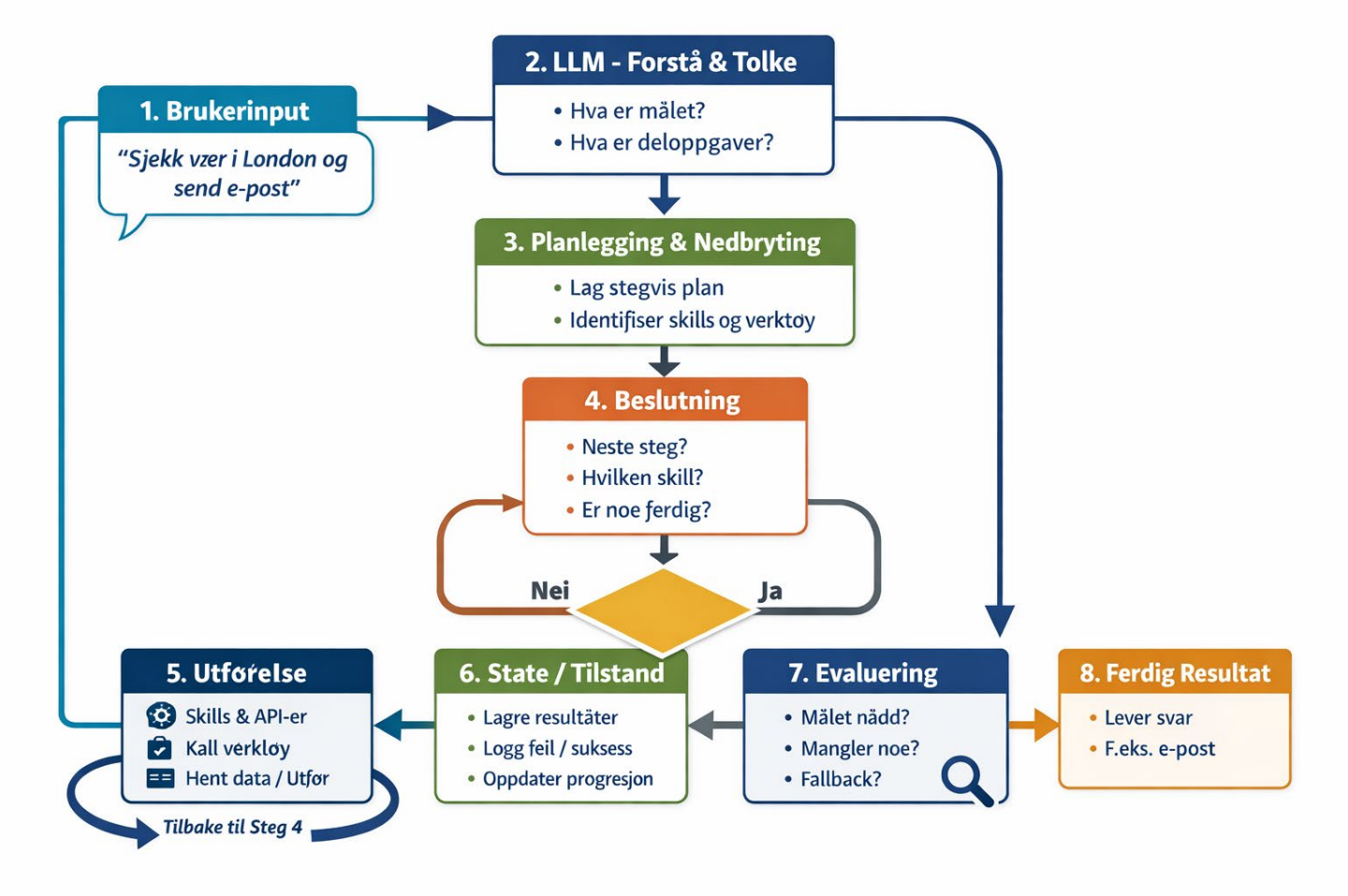
Har kommunen gjennomførende rutiner for å gjenopprette tilgjengelighet til data og systemer ved oppståtte sikkerhetshendelser?

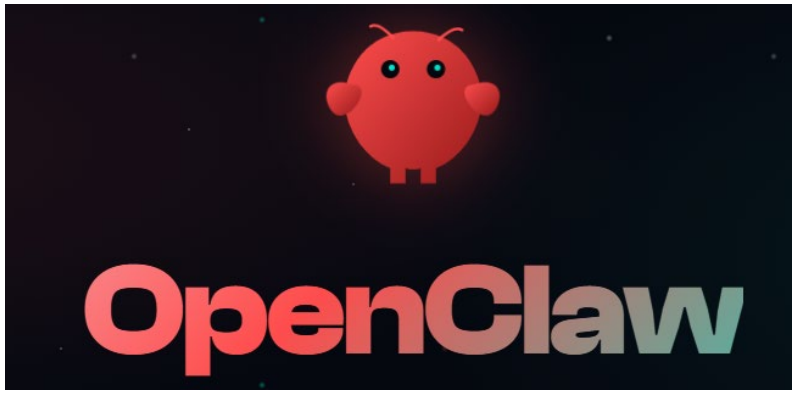
Ja Nei

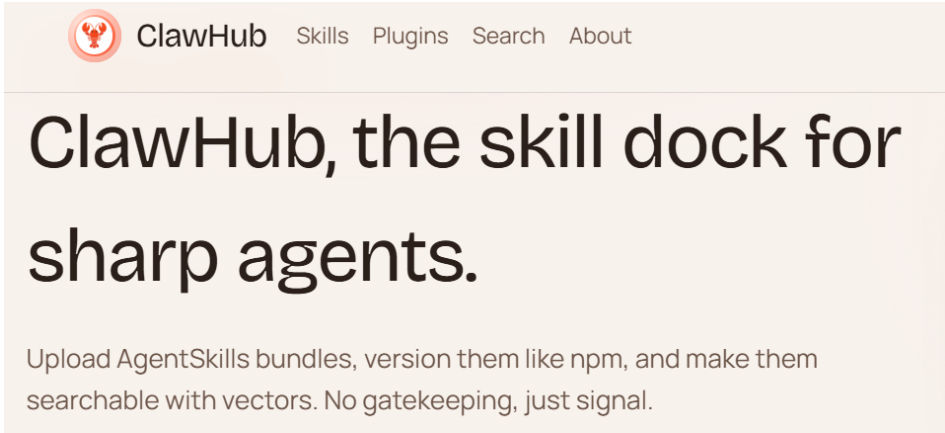
Gjennomfører kommunen jevnlige tester og øvelser knyttet til gjenoppretting av data og systemer?

Ja Nei

Agentisk KI – språkmodeller som faktisk GJØR noe

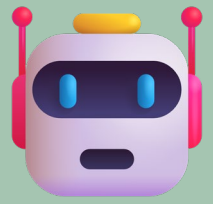






Sara Rossebø
Eirik Gulbrandsen
Datatilsynets seksjon for Teknologi, Sikkerhet og Tilsyn

sara.rossebo@datatilsynet.no
eirik.gulbrandsen@datatilsynet.no



Datatilsynet

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

datatilsynet.no
personvernbloggen.no