Risikoer som bør vurderes i forbindelse med innføring av Microsoft 365 CoPilot

Innholdsfortegnelse

1. Innledning

Dette dokumentet beskriver risikoer som bør vurderes i forbindelse med innføring av Microsoft 365 CoPilot (CoPilot). Dokumentet er utarbeidet av KiNS, og kan brukes som grunnlag for gjennomføring av risikovurderinger i egen virksomhet.

Kommuner som behandlingsansvarlige bør gjennomføre risikovurdering av personopplysningssikkerheten (ROS), jf GDPR artikkel 32. Kommuner bør i tillegg vurdere å gjennomføre vurdering av personvernkonsekvenser (DPIA), jf GDPR artikkel 35. Bruk av CoPilot kan oppfylle flere kriterier som indikerer høy risiko, herunder evaluering eller poengsetting, automatiske beslutninger med rettslig virkning, særlige kategorier personopplysninger, behandling i stor skala, matching eller sammenstilling av datasett og innovativ bruk av teknologi, jf WP29-gruppens kriteriesett.

Risikoene som beskrives i dette dokumentet er i all hovedsak relatert til CoPilot. Kommunene bør vurdere om det er andre risikoer som bør inkluderes i risikovurderingen.

Kommunene bør i tillegg gjennomføre en risikovurdering av selve Microsoft 365. Det bør gjøres i forkant av å risikovurdere CoPilot.

2. Generelt om CoPilot

Microsoft 365 CoPilot er en KI-assistent som er integrert i Microsoft 365-applikasjoner som Word, Excel, Teams og Outlook. CoPilot bruker maskinlæring for å hjelpe brukere med å generere tekst, analysere data, lage presentasjoner og forbedre produktiviteten.

CoPilot er tett integrert med M365 applikasjonene, noe som blant annet innebærer at CoPilot som utgangspunkt har tilgang på brukernes profilinformasjon, samt all informasjon og dokumenter som brukerne selv har tilgang til. Dette inkluderer følgelig også informasjon som kan være av privat karakter, som private e-poster og chatdialoger. Den tette integrasjonen utgjør et markant skille til andre KI-assistenter som for eksempel ChatGPT. Mens CoPilot kan automatisk analysere et dokument i Word, så må brukere manuelt laste opp dokumentet til ChatGPT for analyse. Den brede informasjonstilgangen øker følgelig risikobildet.

3. Formålsvurderinger

Risikobildet ved å ta i bruk CoPilot må sees opp mot av hva CoPilot skal brukes til. Et formål kan være å gjennomføre innsikt og analyse, for eksempel å få bedre innsikt i egne data. Et annet formål kan være å understøtte beslutninger, for eksempel i forbindelse med vedtak eller karaktersetting. Risikobildet relatert til beslutningsstøtte vil typisk være høyere enn ved innsikt og analyse.

4. Risikoer relatert til CoPilot

4.1 Risikoer relatert til samfunn og miljø

Disse risikoene omfatter ressursbruk og miljøpåvirkning. De inkluderer risikoer for høye beregningskostnader og deres påvirkning på miljøet, grunnet ressursintensiv modelltrening og drift.

Risiko	Risikoreduserende tiltak
KI-assistenter er ressurskrevende, både med tanke på modelltrening og drift. Dette kan påvirke kommunes bærekraftsmål eller klimaregnskap negativt.	Innhent informasjon om hvorvidt CoPilot tjenestene drives av fornybar energi. Sikre at Microsoft sine relevante datasentre er karbonnøytrale. Begrens bruken av CoPilot til nødvendige aktiviteter.
Risiko for negative sosiale konsekvenser, inkludert tap av arbeidsplasser eller endringer i arbeidsprosesser på grunn av automatisering. Kan føre til motstand mot KI-teknologi. Kan påvirke kommunens omdømme.	Endringsledelse. Etabler kompetanseutviklingsprogram.

4.2 Risikoer relatert til innføring og forvaltning

Disse risikoene omfatter innføring og bruk av CoPilot. De inkluderer blant annet risikoer for motstand mot endring fra ansatte, utilstrekkelig opplæring og manglende bevissthet om bruk av CoPilot.

Risiko	Risikoreduserende tiltak
Endringsmotstand fra ansatte kan føre til at CoPilot ikke blir brukt, og at formålet med innføringen ikke blir nådd.	Onboard de ansatte i bruken og formålet med CoPilot. Endringsledelse.
Utilstrekkelig opplæring og bevissthet blant ansatte om bruk og risikoer forbundet med CoPilot kan føre til feil bruk av CoPilot.	Gjennomfør opplæringstiltak.
Manglende interne retningslinjer og prosedyrer som regulerer riktig bruk av CoPilot kan føre til feil eller uansvarlig bruk av CoPilot.	Etabler policy for bruk av KI-løsninger. Oppdater sikkerhetsinstruksen.

Kommunen følger ikke opp ansattes bruk av CoPilot, og avdekker følgelig ikke hvorvidt bruken er i henhold til kommunen policy.	Gjennomgå bruksrapporter i Microsoft Purview eller i Viva Insight.
Økt teknologiavhengighet kan redusere ansattes kritiske tenkning og analyseforståelse.	Implementer CoPilot som et støtteverktøy som skal supplere ansattes analysekompetanse.
Nye versjoner av CoPilot («waves») introduserer ny funksjonalitet som ikke blir risikovurdert. Nye versjoner introduserer nye risikoer.	Etabler forvaltning av M365 og CoPilot, herunder en prosess for å vurdere ny funksjonalitet.

5.2 Risikoer relatert til informasjonssikkerhet

Disse risikoene omfatter risikoer relatert til konfidensialitet, integritet og tilgjengelighet, og bør gjennomgås i forbindelse med risikovurderinger av informasjonssikkerheten (ROS).

Risiko	Risikoreduserende tiltak
Tekniske angrep som har til hensikt å omgå de innebygde etiske begrensningene i CoPilot (for eksempel evasion, model extraction, inference attack, prompt injection). Kan føre til uetisk bruk av CoPilot.	Sett opp Insider Risk Management policy i Purview. Gjennomgå aktivitetslogger i Purview.
Trusselaktører kompromitterer ansattes brukerkonto, og får tak i samtalehistorikken. Chathistorikken kan inneholde personopplysninger eller sensitiv informasjon. Brudd på konfidensialitet.	Sikre identitetene med sterk autentisering. Sett opp Conditional Access policies for CoPilot i Entra ID. Sett opp Insider Risk Management policyer i Purview. Lær ansatte hvordan de kan slette egen historikk.
Copilot introduserer nye sårbarheter i det eksisterende IT-miljøet, som kan utnyttes av trusselaktører. Brudd på konfidensialitet, integritet og tilgjengelighet.	Overvåk endepunktene med EDR-løsning. Send telemetri til Defender XDR eller SIEM-løsning. Overvåk og responder.
De ansatte bruker andre KI-tjenester enn CoPilot (for eksempel ChatGPT), som ikke er regulert med sikkerhetstiltak. Brudd på konfidensialitet.	Sett opp policyer i Microsoft Purview som overvåker bruken av andre KI-løsninger. Blokker andre løsninger i brannvegg. Oppdater sikkerhetsinstruksen.

Svak tilgangsstyring i M365 (for eksempel Sharepoint) fører til at CoPilot henter frem informasjon som brukeren i utgangspunktet ikke skal ha tilgang til. Brudd på konfidensialitet.	Gjennomgå tilgangsrettighetene i M365. Sikre «orden i eget hus».
CoPilot er utilgjengelig over tid. Ansatte får ikke brukt tjenesten til å utføre arbeid. Brudd på tilgjengelighet.	Etabler manuelle alternative prosesser for CoPilot-aktiviter.

5.3 Risikoer relatert til datakvalitet, bias og representativitet

Disse risikoene omfatter risikoer relatert til hvordan CoPilot er trent, og risikoer som følger av at datagrunnlaget ikke er representativt for den faktiske bruken. Bør vurderes i forbindelse med vurdering av personvernkonsekvenser (DPIA).

Risiko	Risikoreduserende tiltak
Dataene som er brukt til å trene Copilot er ikke representative for den faktiske bruken eller konteksten. Kan føre til feilaktig analyse og beslutninger. Brudd på prinsippet om riktighet.	Gjennomfør testing av CoPilot. Analyser tilbakemeldingene fra CoPilot. Legg til rette for tilbakemeldinger fra ansatte.
Dataene i M365 tenanten inneholder iboende bias. Kan føre til diskriminerende beslutninger. Brudd på prinsippet om riktighet.	Gjennomgå datagrunnlaget i M365. Slett gammelt og utdatert innhold.
Copilot har tilgang til gamle og utdaterte data i M365. Kan føre til behandling av gamle personopplysninger. Brudd på prinsippet om riktighet.	Søk opp gamle data i Microsoft Purview. Slett gamle og utdaterte data.
CoPilot diskriminerer visse grupper basert på kjønn, rase, seksuell legning, religion osv. Kan føre til juridiske konsekvenser for kommunen, og negative sosiale konsekvenser for de registrerte. Brudd på prinsippet om lovlighet, rettferdighet og åpenhet.	Gjennomfør testing av CoPilot. Analyser tilbakemeldingene fra CoPilot. Legg til rette for tilbakemeldinger fra ansatte.

5.3 Risikoer relatert til personvernprinsippene

Disse risikoene berører krav og plikter kommunen har som behandlingsansvarlig, og bør gjennomgås i forbindelse med vurdering av personvernkonsekvenser (DPIA).

Risiko	Risikoreduserende tiltak
Personopplysninger på avveie, for eksempel at ansatte bruker andre KI-løsninger som ChatGPT. Brudd på prinsippet om integritet og konfidensialitet.	Klassifiser dokumentene i M365 med følsomhetsetiketter. Definer tilgangskontroll på etikettene. Oppdater sikkerhetsinstruksen. Overvåk bruken av CoPilot.
CoPilot analyserer informasjon av privat karakter (for eksempel private e-poster i tenanten, chatdialoger etc). Kan føre til utilsiktet eksponering av personopplysninger. Brudd på prinsippene lovlighet, formålsbegrensning og integritet og konfidensialitet.	Oppdater sikkerhetsinstruksen og presiser hvorvidt privat innhold bør lagres i M365. Vurder egen følsomhetsetikett som kan brukes for å merke privat innhold.
CoPilot analyserer særlige kategorier personopplysninger. Kan føre til utilsiktet eksponering av sensitive personopplysninger. Brudd på prinsippet om integritet og konfidensialitet.	Kartlegg sensitive personopplysninger i M365. Kan gjøres med tilpassede sensitive informasjonstyper i Purview.
CoPilot sammenstiller personopplysninger på tvers av behandlingsaktiviteter, som hver for seg er regulert med ulike formål og rettslige grunnlag (overbehandling og formålsutglidning). Brudd på prinsippene om lovlighet, formålsbegrensning og dataminimering.	Etabler policy for bruk av KI-løsninger, og presiser hvilke aktiviteter CoPilot skal brukes til.
Ledere bruker CoPilot for å overvåke og måle ansattes prestasjoner og atferd (bossware). Kan føre til ulovlig overvåking av ansatte. Brudd på prinsippene om lovlighet og integritet og konfidensialitet.	Etabler policy for bruk av KI-løsninger, og presiser hvilke aktiviteter CoPilot skal brukes til.
Ansatte med tilgang til eDiscovery i Purview kan eksportere ansattes samtaledialoger. Brudd på integritet og	Sikre at personer som skal ha tilgang til eDiscovery må gjennomgå en godkjenningsprosess (for eksempel

konfidensialitet. Brudd på e-postboksforskriften.	elevering gjennom Microsoft Entra ID PIM).
CoPilot samtaledialoger som inneholder personopplysninger lagres over tid. Brudd på prinsippet om lagringsbegrensning.	Lær ansatte hvordan de kan slette egen historikk.

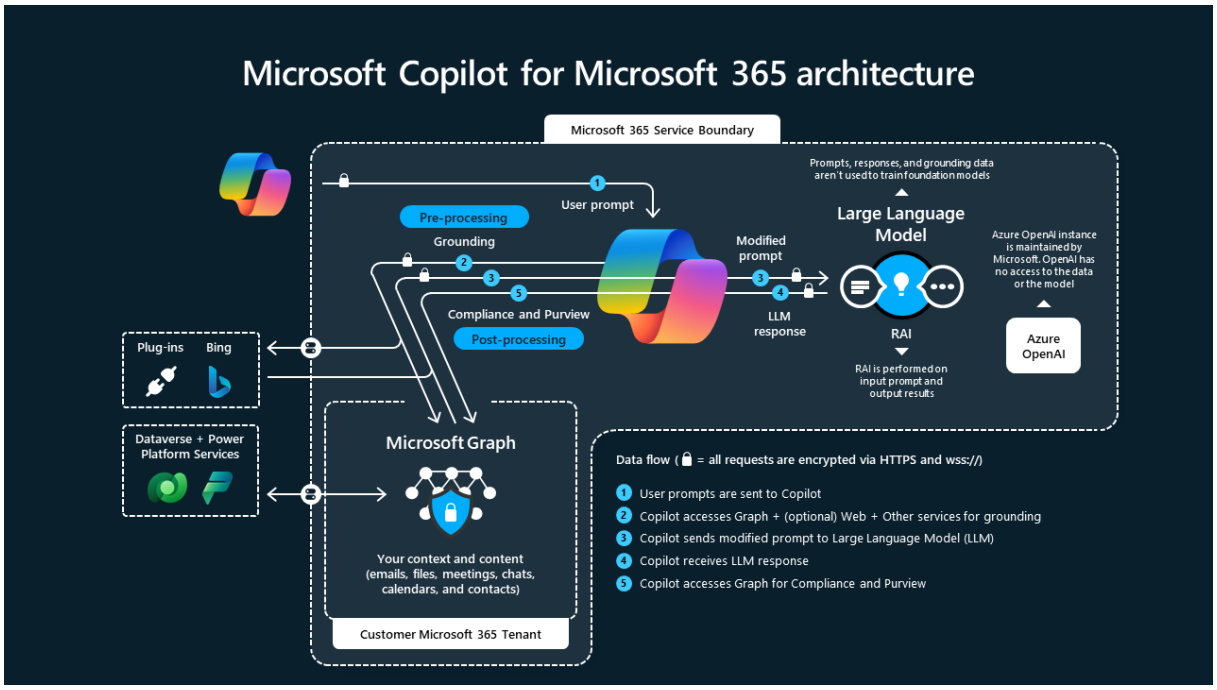
5.4 Risikoer relatert til de registrertes rettigheter

Disse risikoene berører de registrertes rettigheter og friheter, og bør gjennomgås i forbindelse med vurdering av personvernkonsekvenser (DPIA).

Risiko	Mulig konsekvens
Manglende samsvar mellom de registrertes rettigheter og den behandlingsansvarliges plikter etter personvernregelverket, for eksempel rett til innsyn, sletting osv. Kommunen som behandlingsansvarlig etterkommer ikke de registrertes rettigheter.	Etabler policy for bruk av KI-løsninger. Etabler relevante prosedyrer.
CoPilot brukes for å gjennomføre automatiserte beslutninger. De registrerte har rett til å reservere seg mot automatiserte beslutninger. Brudd på GDPR artikkel 22 (automatiserte individuelle avgjørelser).	Etabler policy for bruk av KI-løsninger, og presiser hvilke aktiviteter CoPilot skal brukes til.
Manglende forståelse av hvordan Copilot kommer fram til det genererte innholdet, noe som kan vanskeliggjøre verifikasjon og redusere tillit. Manglende innsikt i CoPilot sine beslutningsprosesser (black box problemet) kan føre til manglende forutsigbarhet.	Etabler policy for bruk av KI-løsninger, og presiser hvilke aktiviteter CoPilot skal brukes til. Unngå å bruke CoPilot til beslutningsprosesser.

5. Regulering av CoPilot med Purview

Den tekniske arkitekturen i CoPilot kan illustreres slik:



Illustrasjonen viser at det er mulig å regulere bruken av CoPilot med Microsoft Purview (se punkt 5 i illustrasjonen). Generelt sett bør følgende risikoreduserende tiltak gjennomføres i Purview:

1. Få innsikt i dataen som ligger i M365 (for eksempel dokumenter i Sharepoint og e-poster i Exchange). Siden dataen utgjør CoPilot sitt kunnskapsgrunnlag, så bør gamle eller utdaterte data slettes. Bruk Content Explorer og Content Search for å få oversikt over all informasjon i M365 tenanten.

2. Klassifiser e-poster og dokumenter. Publiser følsomhetsetiketter til ansatte, og etabler prosedyrer for at ansatte merker e-poster og dokumenter med følsomhetsetikettene. Vurder om CoPilot skal ha tilgang til e-poster og dokumenter merket konfidensielt og/eller privat.

3. Få innsikt i bruken av CoPilot og andre KI-assistenter. Purview gir rapporter på faktisk bruk av CoPilot, og hvorvidt ansatte bruker andre KI-assistenter som ChatGPT. Merk at dette forutsetter korrekt lisensiering og noe teknisk tilrettelegging på endepunktene.

6. Grunnleggende prinsipper for utvikling av KI-tjenester

Microsoft har definert følgende seks prinsipper som de bruker i forbindelse med utvikling av KI-tjenester. Prinsippene skal være fulgt i utformingen av Microsoft 365 CoPilot, og kan brukes for å utlede ytterligere risikoer.

1. Rettferdighet: KI-systemer bør behandle alle mennesker rettferdig.
   Det er viktig at KI-systemer ikke diskriminerer eller viser bias mot noen basert på kjønn, rase, seksuell orientering, religion eller andre karakteristika. Rettferdighet innebærer at alle beslutninger tatt av KI er objektive og ikke skadelige for noen spesifikke grupper eller individer.

2. Pålitelighet og sikkerhet: KI-systemer bør operere pålitelig og sikkert.
   KI-systemer må fungere som de er designet for og kunne håndtere nye situasjoner trygt. Dette innebærer grundig testing og validering under forskjellige forhold for å sikre at systemet kan håndtere ekstreme tilfeller og motstå manipulasjon.

3. Personvern og sikkerhet: KI-systemer bør være sikre og respektere personvernet.
   Beskyttelse av persondata er avgjørende. KI-systemer må sikre at dataene de bruker er beskyttet mot uautorisert tilgang, og at bruken av disse dataene ikke kompromitterer individers personvern.

4. Inkludering: KI-systemer bør styrke og inkludere alle mennesker.
   Inkluderende designpraksiser bidrar til å forstå og adressere potensielle barrierer som kan ekskludere mennesker. Dette inkluderer bruk av teknologier som tale-til-tekst og tekst-til-tale for å støtte personer med nedsatt funksjonsevne.

5. Transparens: KI-systemer bør være forståelige.
   Transparens handler om å sikre at informasjon om KI-systemer og deres beslutningsprosesser er tilgjengelig og forståelig for brukerne. Dette inkluderer å forklare hvilke data og algoritmer som brukes, og hvordan modellene fungerer.

6. Ansvarlighet: mennesker bør være ansvarlige for KI-systemer.
   Det bør være klart hvem som er ansvarlig for KI-systemenes handlinger og beslutninger. Dette innebærer å ha klare ansvarsstrukturer og mekanismer på plass for å sikre at AI brukes på en etisk og ansvarlig måte.

7. Hovedpunkter fra rapporten «CoPilot med personvernbriller på»

NTNU sin sluttrapport for vurdering av CoPilot omfatter følgende hovedpunkter. De kan også brukes for å utlede ytterligere risikoer.

1. M365 CoPilot forutsetter at virksomhetsdata ligger i Microsofts skyløsning.
2. Sørg for orden i eget hus.
3. Identifiser og begrens hva copiloten skal brukes til.
4. Vurder det rettslige grunnlaget.
5. Vurder personvernkonsekvenser.
6. Vil bruken være i strid med e-postforskriften?
7. Bruk av språkmodeller krever kompetanse og bevissthet.
8. Vurder alternative løsninger.
9. Gjør innføringen i små og kontrollerte steg.

Referanser

- NIST AI Risk Management Framework (AI RMF 1.0): https://www.nist.gov/itl/ai-risk-management-framework

- Microsoft 365 CoPilot documentation: https://learn.microsoft.com/en-us/copilot/microsoft-365/

- CoPilot for Microsoft 365 i offentlig sektor: https://www.ntnu.no/adm/it/copilot

- CoPilot med personvernbriller på: https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/ferdige-prosjekter-og-rapporter/ntnu-sluttrapport-copilot-med-personvernbriller-pa/

- CoPilot sandkasseprosjekt hos Datatilsynet: https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/ferdige-prosjekter-og-rapporter/ntnu-sluttrapport-copilot-med-personvernbriller-pa/hvordan-kan-m365-copilot-forstas-i-lys-av-personvernregelverket/