



Kunstig intelligens i digitale læringsressurser og fagsystem

En veileder fra SkoleSec-prosjektet

Versjon	Dato	Kommentar	Endret av
1.0	15.01.2025	Publisert første versjon	SH
1.1	16.01.2025	Mindre språklige endringer	AJ

INNHold

1	Innledning	4
2	Bakgrunn	6
3	Omfang og oppbygning	7
3.1	Omfang og målgruppe	7
3.2	Roller	7
4	Tjenester for kunstig intelligens	9
4.1	Kunstig intelligens i digitale læringsressurser og fagsystem	9
4.2	Maskinlæringsmodeller – byggeklosser i ki-systemer	9
4.3	Synkron og asynkron respons	12
4.4	Rollen til skytjenesteplattformer	13
5	Personopplysningsvern i kontekst av opplæring	15
5.1	Behandlingsgrunnlag	15
5.2	Personvernrisiko	15
6	Risiko for behandling av personopplysninger ved elever og ansattes bruk av kunstig intelligens i opplæringen	17
6.1	Manglende åpenhet i maskinlæringsmodeller og tjenester for kunstig intelligens	18
6.1.1	Risiko og konsekvens ved manglende åpenhet	26
6.2	Fra beslutningsstøtte til beslutning	26
6.2.1	Artikkel 22: Automatiserte beslutninger og profilering	27
6.2.2	KI-forordningen: Høyrisikoformål og kompetansekrav	30
6.2.3	Risiko og konsekvens ved kunstig intelligens som beslutningsstøtte	32
6.3	Viderebehandling uforenlig med opprinnelig formål og rolleimplikasjoner	33
6.3.1	Risiko og konsekvens ved viderebehandling	35
6.4	Integrasjon med eksterne datakilder eller prosesser	36
6.4.1	Risiko og konsekvens ved integrasjon med eksterne datakilder eller prosesser ...	37
6.5	Reidentifisering av pseudonymiserte eller anonymiserte opplysninger	38
6.5.1	Risiko og konsekvens ved reidentifisering	40
6.6	Misvisende, villedende eller skadelig innhold	40
6.6.1	Risiko og konsekvens ved misvisende, villedende eller skadelig innhold	42
7	Behandlingsaktiviteter	43
8	Anbefalte tiltak for skoleeier	45
8.1	Lovlighet	45

8.2	Rettferdighet.....	47
8.3	Åpenhet	47
8.4	Formålsbegrensning	48
8.5	Dataminimering.....	48
8.6	Riktighet.....	49
8.7	Lagringsbegrensning	50
8.8	Integritet og konfidensialitet.....	51
8.9	Ansvarlighet og behandlingsansvarliges videre vurderinger	51
9	Kilder	53
10	Vedlegg.....	57

1 INNLEDNING

Opplæringen har som formål å åpne dører mot verden og fremtiden, og den skal gjøre elever og lærlinger i stand til å tenke kritisk og handle etisk og miljøbevisst, jf. opplæringslova § 1-3.

Teknologi er en vesentlig og integrert del av den verden de går i møte. I den moderne skolen, der eleven både har rett og ansvar til å ta aktiv del i egen læring, må teknologi brukes i samhandling med andre elever, og med læreren, for å oppnå målene i læreplanverket. Å innrette skolen og elevenes læring mot reelle problemstillinger i samfunnet, innebærer også i mange tilfeller å ta i bruk et bredt utvalg av læringsressurser, også digitale. Disse vil i mange tilfeller behandle personopplysninger og i økende grad inneholde kunstig intelligens. Det stiller særlige krav til skolen og skoleeier for å sørge for at utvikling og bruk er i tråd med formål med bruk, og formål i lovgivning.

Kunstig intelligens, forkortet KI, er et paraplybegrep for flere ulike former for algoritmisk datateknologi utviklet gjennom maskinlæringsteknikker, slik som dyplæring. Dette er en type prediksjonssystemer, trent til å justere og tilpasse respons basert på ulike inndata, slik at de tilsynelatende fremstår som intelligente. Kunstig intelligens forbindes ofte i dag med generativ kunstig intelligens, som er systemer som kan skape ulike former for medieinnhold, slik som tekst, bilde, lyd og video. KI som implementeres i digitale læringsressurser og pedagogiske fagsystemer, utvider ytteevne til disse ressursene og gjør at de kan benyttes til å oppfylle et bredt spekter med ulike formål. Det inkluderer potensielt nye formål utover det opprinnelig tiltenkte, og der det ikke foreligger et bredt kunnskapsgrunnlag. Der teknologien implementeres i digitale læringsressurser i skolen, og anvendes for pedagogiske og skoleadministrative formål, kan dette medføre risikoer og mulige negative konsekvenser for eleven eller lærlingens rettigheter og friheter. Både teknologiens muligheter og begrensninger, samt formål og metoder for implementering i praksis, bidrar til å skape nye risikoer. Det er vanskelig for personer som bruker systemer med kunstig intelligens å overprøve eller stille spørsmål ved en respons som virker plausibel. Dette kan bidra til en forventning om at responsen til et KI-system generaliserer til oppgaver for andre formål. Behandlinger med generativ KI er i prinsippet ikke-deterministisk¹ og relativt små justeringer til inndata kan gi store utslag i utdata. Det kan da være vanskelig å si sikkert at et KI-system har gjennomført en behandling i tråd med formålet. For formål som krever høy grad av validitet² og reliabilitet³, så kan det være vanskelig å ivareta personvernprinsippene; særlig åpenhet, dataminimering og formålsbegrensning. Risikoene for de registrertes rettigheter og friheter skal vurderes i lys av dette, med et særskilt krav til en personvernkonsekvensvurdering jf. artikkel 35, der behandlingene utgjør høy risiko for elever og lærlinger.

Ny forordning om kunstig intelligens (Regjeringen, 2024) vil stille særskilte krav til bruken av kunstig intelligens til enkelte opplærings- og utdanningsformål, som er klassifisert som *høyrisiko*.

¹ At et KI-system er ikke-deterministisk betyr at vi ikke kan kjenne utfallet/responsen til systemet ikke er kjent på forhånd eller årsaksbestemt.

² Validitet handler om hvorvidt et resultat gir grunnlag for å trekke holdbare slutninger om det formålet resultatet skal belyse.

³ Reliabilitet handler om hvorvidt man får samme resultat ved å gjenta noe gjentatte ganger.

Denne veilederen peker frem til forpliktelsene som ligger i KI-forordningen, som krever overensstemmelse med både GDPR og ePrivacy-direktivet, og som harmoniserer kravene om behandling av personopplysninger ved bruk av kunstig intelligens. Denne veilederen vil slik sett måtte utvikles videre i takt med utvikling av teknologi, praksis og lovverk, slik at skoleeier har nødvendig støtte til å kunne tilby læringsressurser med kunstig intelligens, i tråd med læreplan.

Til tross for at det må forventes et vesentlig arbeid med å ivareta ansvarlighet ved bruk av digitale læringsressurser med kunstig intelligens og de nødvendige behandlingene til tider kan være vanskelig å dokumentere, vurdere og forklare, så taler ikke det i seg selv for å velge bort disse ressursene. Det taler derimot for en refleksiv tilnærming til både pedagogisk- og skoleadministrativ praksis, og at disse foregår i et slikt samspill at ansvarlighetsprinsippet ivaretas. Skolens samfunnsoppdrag er å møte samfunnsutvikling og utvikle nødvendige kompetanser hos elev, lærling, praksiskandidat og lære kandidat til videre utdanning og arbeid, og demokratisk medborgerskap. Kunstig intelligens, som både teknologi og teknologisk prosess, er tett forbundet med en hensiktsmessig bruk av digitale læringsressurser i lærerens pedagogiske og fagdidaktiske praksis. Artikkel 4 av KI-forordningen stiller krav om en kontekstspesifikk kompetanse, som blant annet hensyntar personene eller gruppene som KI systemet skal brukes på. Rammeverket for lærerens profesjonsfaglige digitale kompetanse (PfDK) kan til en viss grad sies å ivareta en slik kontekstspesifikk tilnærming og vil kunne bidra til å ivareta elevens rettigheter og friheter. Dette må operasjonaliseres nærmere i en fremtidig veileder for lærers implementering i pedagogisk praksis.

Denne vurderingen skal danne grunnlag for kommuner og fylkeskommuners ivaretagelse av behandlingsansvaret slik det er definert i personopplysningsloven (GDPR, EU 2016/709), med formål om å utføre oppgaver etter opplæringslova med forskrifter. Implementeringstakten for KI-systemer som brukes til å oppfylle disse formålene tilsier at arbeidet med utvikling av et forvaltningsrammeverk for KI må igangsettes. Denne veilederen og DPIA for kunstig intelligens i digitale læringsressurser og fagsystem skal hjelpe kommuner og fylkeskommuner å utvikle og ta i bruk denne teknologien på måter som ivaretar både individet og samfunnets interesser.

2 BAKGRUNN

Et økende antall digitale læringsressurser og pedagogiske fagsystem integrerer funksjonalitet for læringsanalyse og vurderingsstøtte med behandlinger som utføres av kunstig intelligens. Integrering av denne type funksjonalitet i tjenester tiltenkt opplæring, foregår i sammenheng med utviklingen av et stort antall netjtjenester med generativ KI som kjerneteknologi. Særlig aktuelt for tiden er chatbot-tjenester som ChatGPT, Character.ai, JanitorAI og MyAI, samt KI-basert søk, som Perplexity og YOU, som alle bygger på generative språkmodeller. Tilgang til den samme språkteknologien kan kjøpes som integrasjoner, og relativt enkelt gjøres til en del av digitale læringsressurser og fagsystem. Denne type kunstig intelligens er en relativt ny teknologi og kan ha betydelige tilknyttede risikoer, når den anvendes i opplæringen til en elev.

Hvis det er sannsynlig at en behandling av personopplysninger vil medføre høy risiko for de registrertes rettigheter og friheter skal det gjennomføres en vurdering av personvernkonsekvenser (DPIA), jf. personvernforordningen art. 35 nr. 1. Datatilsynet har utarbeidet en liste over hvilke behandlingsaktiviteter som alltid krever en vurdering av personvernkonsekvenser, og det er flere punkter som er relevante for bruk av KI-systemer i skolen:

- Behandling av personopplysninger for å evaluere læring, mestring og trivsel i skoler eller barnehager. Dette inkluderer alle utdanningsnivåer, fra barne- og ungdomsskole, videregående skoler og høyere utdanning.
- Behandling av personopplysninger med innovativ teknologi i følge med minst ett annet kriterium.
- Innovativ teknologi og stort omfang sårbare registrerte.

Kunstig intelligens, benyttet enten direkte eller indirekte i en behandling av personopplysninger, har potensiale til å medføre konsekvenser for blant annet retten til vern av personopplysninger, av privatliv og egen kommunikasjon, ikke-diskriminering, samt rett til tanke- og ytringsfrihet jf. fortalepunkt 4 i personvernforordningen/GDPR. Implementering av ny teknologi i stort omfang og til behandling av personopplysninger nært knyttet til noens personlige forhold jf. §13 forvaltningsloven, kan føre til høy risiko.

KS, ved SkoleSec-prosjektet, har satt ned en arbeidsgruppe for å vurdere personvernkonsekvenser for bruk av kunstig intelligens i digitale læringsressurser og fagsystem (DPIA). Arbeidsgruppen har bestått av representanter fra Bergen kommune, Kongsvinger kommune, Oslo kommune, Telemark fylkeskommune og Vestfold fylkeskommune.

Arbeidsgruppen vurderer at det må gjennomføres en DPIA for bruk av denne typen kunstig intelligens i digitale læringsressurser og fagsystem.

3 OMFANG OG OPPBYGNING

3.1 OMFANG OG MÅLGRUPPE

Denne vurderingen består av en overordnet veileder (dokumentet du leser nå), samt en «case-DPIA». Temaet er behandling av personopplysninger med skybasert kunstig intelligens, implementert gjennom et applikasjonsprogrammeringsgrensesnitt (API), og brukt av elever i en digital læringsressurs og av lærere i et pedagogisk fagsystem. Denne veilederen drøfter primært språkmodeller og generativ kunstig intelligens, da dette er den type maskinlæringsmodeller som skaper responser som er både mest synlig for sluttbrukeren og enklest å nyttiggjøre seg av. Det er samtidig også en type kunstig intelligens for generelle formål, noe som innebærer at teknologien relativt enkelt og uten vesentlige hinder kan tas i bruk til formål definert av sluttbrukeren, i motsetning til et system med avgrenset funksjonalitet og kapabilitet. Det er foretatt tester av ulike komponenter, for å bedre forstå deres design og kapabiliteter. Det er studert både hvordan tjenestene er innrettet mot lærer og mot elev, og konseptuelt sett, hvordan behandling av personopplysninger vil forekomme i disse sammenhenger (fig.1 s.7).

I veilederen tar arbeidsgruppen for seg implementering og bruk av kunstig intelligens og problemstillinger knyttet til behandling av personopplysninger med kunstig intelligens. Formålet her er å identifisere risikoer og konsekvenser ved bruk av læringsressurser med KI generelt, og foreslå noen overordnede tiltak som kan være relevante i de fleste tilfeller. I «case-DPIAen» har arbeidsgruppen vurdert behandlingsaktiviteter ved bruk av KI, eksemplifisert ved en konkret læringsressurs, LearnLab. Formålet er å adressere risiko og tiltak ved en skybasert KI, som er implementert gjennom API-er. Ettersom feltet er i endring og det er flere KI-tjenester som vurderes, så skal veilederen holdes oppdatert og fungere som en rettesnor for videre vurderinger av risiko og konsekvenser i enkelte prosesser og lærings-ressurser.

Målgruppen for denne vurderingen er roller som representerer skoleeier som behandlingsansvarlig i vurdering av digitale tjenester som behandler personopplysninger, og de som utvikler og gjennomfører kompetansehevende tiltak. Det legges til grunn at de som sammen skal anvende veilederen har en grunnleggende forståelse av personvernforordningen (GDPR), opplæringslova med forskrifter, pedagogisk og skoleadministrativ praksis, samt digitale teknologier og informasjonssystem; slik som skytjenester og applikasjoner. Veilederen anbefales brukt sammen med andre ressurser på skolesec.no, samt skoleeiers styrende dokumenter. Det er også viktig å se til blant annet personopplysningsloven, EDPB-veiledere om GDPR, opplæringslova med forskrifter (opplæringsforskrifta og læreplanverket), teknisk dokumentasjon om aktuelle systemer, samt beskrivelser av og rutiner for oppgavene som skal utføres.

3.2 ROLLER

De registrerte er i denne sammenhengen elever, lærlinger og ansatte i skolen.

Behandlingsansvarlig er den som bestemmer formål og midler ved en behandling. Denne vurderingen legger til grunn at det er kommune og fylkeskommune som skoleeier som har

ansvaret for opplæringen, jf. opplæringslova §§ 28-1, 28-2, og dermed ansvar og myndighet til å definere formål og midler som anvendes for å oppfylle retten til opplæring og utdanning. Det er dermed skoleeier som behandlingsansvarlig sitt ansvar å vurdere om behandlinger som igangsettes ivaretar lovprinsippene for behandling av personopplysninger, jf. artikkel 5 i GDPR.

Databehandler er den som utfører en behandling på vegne av behandlingsansvarlig, etter dens instruks.

4 TJENESTER FOR KUNSTIG INTELLIGENS

4.1 KUNSTIG INTELLIGENS I DIGITALE LÆRINGSRESSURSER OG FAGSYSTEM

Kunstig intelligens er en samlebetegnelse for maskinlæringsmodeller trent på ulike eksempler, med formål om å generalisere til ulike oppgaver. Dette kan for eksempel være å klassifisere bilder, tekst og andre data, eller å skape nye data basert på eksempler. I digitale læringsressurser og fagsystem kan teknologien brukes til å analysere og klassifisere elevarbeid (adaptivitet), eller til å utvikle KI-assistenter for lærere og elever.

Det er ulike måter å implementere kunstig intelligens i digitale læringsressurser og fagsystem. Det vanligste er å knytte en slik tjeneste til en maskinlæringsmodell i et tredjeparts datasenter. Noen av grunnene til at dette er at:

- maskinlæringsmodeller i mange tilfeller krever spesialisert infrastruktur med høy ytelse og kapasitet, for å utføre såkalt «inferens» (prediksjon).
- kostnadene blir lavere ved å tilby teknologien i stor skala.
- det gir skytilbydere et konkurransefortrinn å tilby modellene med høyest yteevne.

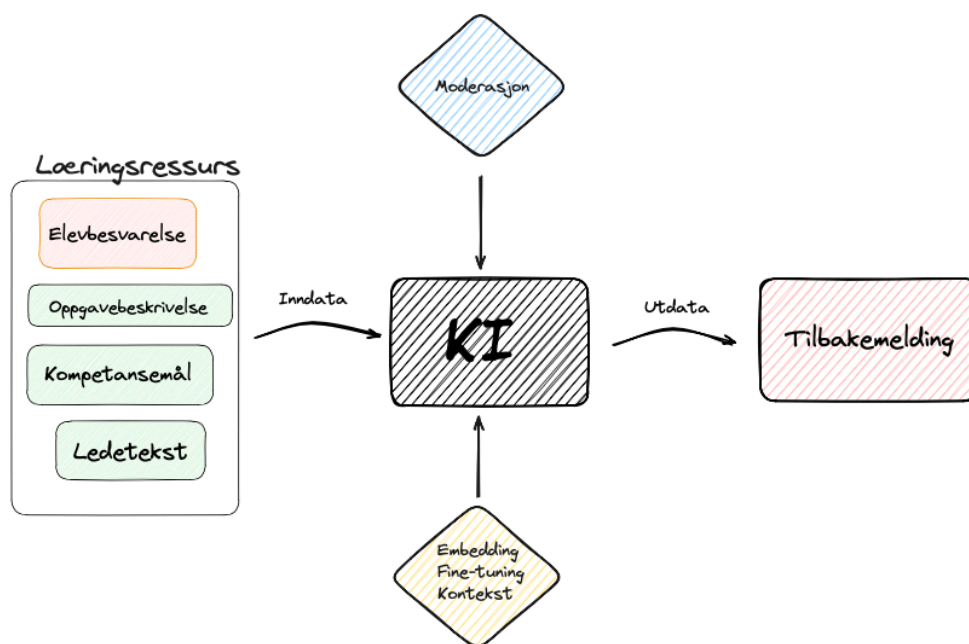
Implementering og bruk av kunstig intelligens tilgjengeliggjort av tredjeparts skytilbydere bringer med seg fordeler for utviklere som ønsker å tilby funksjonaliteten i egne tjenester. Samtidig medfører dette en endring i risikofaktorene for behandlingsansvarlig og den registrerte. Risikoene er i hovedsak knyttet til bruken av skybaserte endepunkt. Dette fordi data må overføres og behandles i et miljø som introduserer flere problemstillinger knyttet til blant annet kryptering og krypteringsnøkkelhåndtering, lagring og innsyn. I tillegg innebærer dette som oftest ytterligere databehandlere, der den behandlingsansvarlige må påse at det ikke uten grunnlag viderebehandles personopplysninger og virksomhetsdata, eller at det forekommer overføringer utenfor EØS-området, som ikke er tilstrekkelig vurdert.

4.2 MASKINLÆRINGSMODELLER – BYGGEKLOSSER I KI-SYSTEMER

En maskinlæringsmodell er én eller flere matematiske funksjoner (vektorer) som er kodet i et program, basert på eksempler gitt under trening (Kelleher, 2019). Dette er det som betegnes som modellens parametere og som ofte brukes som en enhet for å indikere størrelsen på modellen (Den faktiske størrelsen når modellen tas i bruk er avhengig av optimaliseringer; det som kalles «quantization»). En modell kan brukes som den er eller videretrenes for å bedre tilpasses enkelte formål. En modell kan brukes direkte av en person som sender instruksene direkte til modellen, eller gjennom et system som er plassert mellom brukeren og modellen. En modell kan brukes til å trekke ut og fortolke vektene i en annen modell, og til å beregne egne verdier for å skape en respons for enten en bruker eller en annen modell. Tjenester som bruker modeller på denne måten kalles agentiske (Park et al, 2023, Wang et al, 2023). Prosessen med å beregne respons kalles «inferens», og er der modellen instrueres til «å utlede nye setninger fra forutgående» (Russel og Norvig, 2010, s. 235, egen oversettelse).

Opplæringssektoren kjøper i dag typisk tilgang til tjenester med kunstig intelligens eller maskinlæringsmodeller som er ferdig trent og som kan benyttes til å utføre et bredt sett med oppgaver. Slike tjenester eller modeller betegnes som kunstig intelligens til generelle formål, eller general-purpose AI system. I KI-forordningen er dette definert som KI-systemer som viser «signifikant generaliserbarhet og som er kompetent til å utføre et bredt sett med distinkte oppgaver uavhengig av måten modellen er tilgjengeliggjort i markedet og som kan integreres i et utvalg av systemer og applikasjoner» (artikkel 3-63, KI-forordningen, EU 2024/1689, egen oversettelse). For eksempel kan en språkmodell, en type KI til generelle formål, brukes både med formålet om å lage et utkast til en tekst med et gitt tema, men kan også brukes til å vurdere kriterier ved en tekst – to ulike formål. Risikoene ved at maskinlæringsmodeller kan brukes til ulike formål diskuteres nærmere i kapittel 6.1 til 6.3. Det er også mulig å videretrene modellene, noe som diskuteres ytterligere i kapittel 6.4 og 6.6.

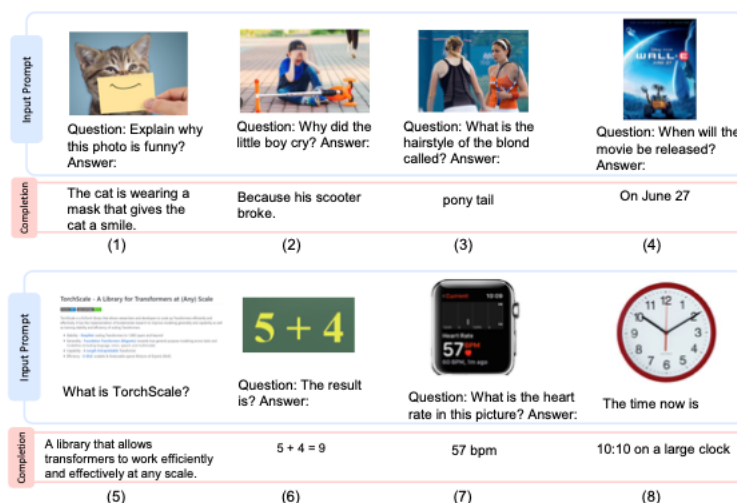
Det at maskinlæringsmodeller kan konfigureres og kombineres til KI-systemer med ulike formål, også av sluttbrukere, stiller særskilte krav til skoleeiers vurderinger av implementeringen av kunstig intelligens. Behandlinger i KI-systemene og prosessene de understøtter må samtidig også sees i sammenheng med de skytjenestene som brukes, og hvordan ulike data behandles i overføring fra klient til server og tilbake, samt eventuelle viderebehandlinger langs denne stien. Arbeidsgruppen har utarbeidet en konseptuell modell for behandling av inn- og utdata i kontekst av en læringsressurs eller pedagogisk fagsystem (figur 1).



Figur 1 - Konseptskisse KI-arkitektur i digital læringsressurs

Modellforklaring (figur 1)	
Læringsressurs	Innrammingen der elev eller lærer definerer inndata.
Inndata	Samlet instruks til KI.
KI	«Den sorte boksen» som inkluderer modellens parametere, vektorer og bias fra trening.
Moderasjon	Filtrering av hva språkmodellen får lov til å skape av innhold.
Embedding	Matematiske representasjoner av innhold for å muliggjøre søk og sammenlikninger.
Fine-tuning	Ulike teknikker for videretrening av modell.
Kontekst	Metainformasjon som instruerer modellen i hvordan den skal respondere på inndata.
Utdata	Utledet respons som utgjør tilbakemelding til bruker eller annet system.
Tilbakemelding	Multimodalt innhold som lærer eller elev nyttiggjør videre i pedagogisk arbeid eller læring.

Læringsressurser og fagsystem med kunstig intelligens kan få kapabiliteter til å tolke bilder, video, lyd og annet medieinnhold, og så sette disse i en kontekst, gjennom å implementere modeller med multimodale kapabiliteter (Chai et al, 2021, Huang et al, 2023). F.eks. å tolke emosjoner i et bilde eller lydklipp. Kunstig intelligens kan brukes til å klassifisere både ikke-strukturerte data og ikke-merkede datasett (Ng, 2011). Kunstig intelligens som både kan tolke og skape ulike typer medieinnhold, kalles multimodale modeller. To typer er blant annet «diffusion»-modeller (Yang et al, 2023) og «transformer»-modeller (Vaswani et al, 2017). Modeller lærer å skape nytt innhold basert på eksempler gitt i trening (Gregor, 2014), noe som også medfører at modellene kan lære å reprodusere eksemplene og eventuell bias representert i disse, uten nødvendigvis å inneha mekanismer for å kunne korrigere utdata (Bender et al, 2021).

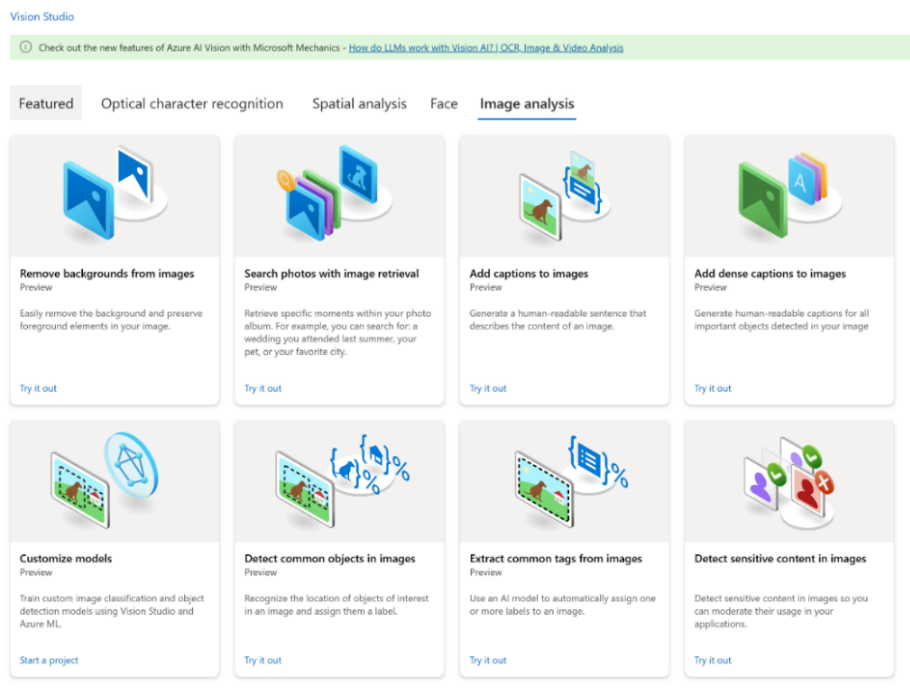


Figur 2 – Multimodal instruks, analyse og respons (Huang et al, 2023)

Maskinlæringsmodeller består av ulike algoritmer eller nevrale nettverk, trent på spesifikke former for innhold. Disse faktorene har betydelig innvirkning på hvordan inndata behandles og

hvordan utdata beregnes; altså hvordan elever og læreres brukerskapt innhold behandles, og hvilken respons de får (jf. figur 1). Åpne språkmodeller tilgjengeliggjør som hovedregel vektene (og i noen tilfeller treningsdata), mens lukkede modeller, slik som OpenAI sine GPT-modeller og Google sine Gemini-modeller, ikke gjør det. I de fleste tilfeller regnes det som brudd på avtalevilkårene å forsøke å trekke ut modellvektene. Dette forholdet mellom åpne og lukkede modeller er vesentlig når personvernprinsippene om åpenhet, riktighet og ansvarlighet vurderes.

Eksempler på bruk av maskinell multimodal analyse, til å identifisere, klassifisere og analysere bilde, video, lyd og tekst:



Figur 2 - Oversikt over kapabiliteter for bildeanalyse i Vision Studio (Schroeder, 2023)

4.3 SYNKRON OG ASYNKRON RESPONS

Interaksjon med ulike varianter av kunstig intelligens, slik som samtaleroboter drevet av språkmodeller, kan være både synkron og asynkron (Ramjee et al, 2024), noe som er en viktig distinksjon for å ramme inn og beskrive ulike bruksscenarier. Der en person kommuniserer med for eksempel en språkmodell, så vil den i interaksjonen i all hovedsak være umiddelbar og i sanntid – altså synkron respons. Der et system med kunstig intelligens brukes til å analysere data og utforme responser etter en tidsplan, så vil det være noe forsinkelse eller planmessighet – altså asynkron respons. Disse bruksscenariene kan være vesentlige i hvordan digitale læringsressurser og pedagogiske fagsystem med KI brukes i undervisning og læring. Dette fordi det kan gi ulike muligheter og utfordringer i forhold til å kvalitetssikre inn- og utdata fra systemet, og hvordan disse tas i bruk.

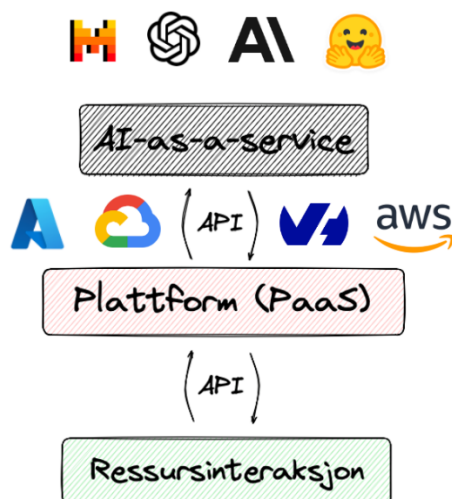
Dersom systemet brukes som beslutningsstøtte, må det vurderes om personen som er ansvarlig for KI-systemet har nødvendig kapasitet, autoritet og kompetanse til å utøve sin rolle i forhold til å vurdere og overprøve både synkron og asynkron respons, og eventuelt hvilke rutiner som må implementeres for å støtte den ansvarlige. Under KI-forordningen vil det stilles krav til automatisk logging av hendelser i enkelte systemer med kunstig intelligens jf. krav til høyrisikosystemer i artikkel 12 av AIA:

- *For high-risk AI systems referred to in point 1 (a) of Annex III, the logging capabilities shall provide, at a minimum:*
 - o *(a) recording of the period of each use of the system (start date and time and end date and time of each use);*
 - o *(b) the reference database against which input data has been checked by the system;*
 - o *(c) the input data for which the search has led to a match;*
 - o *(d) the identification of the natural persons involved in the verification of the results, as referred to in Article 14(5).*

4.4 ROLLEN TIL SKYTJENESTEPLATTFORMER

Plattformene, rammeverkene og integrasjonene som tilgjengeliggjør maskinlæringsmodeller spiller en viktig rolle i tilgjengeliggjøringen av tjenester med kunstig intelligens. Gjennom å integrere en læringsressurs eller et fagsystem med en skyplattform som tilbyr KI-teknologi, kan den relativt enkelt tilbys som funksjonalitet for sluttbrukeren. Forretningsmodellen kan betegnes som «KI-som-tjeneste», ettersom skytjenestene teknologien leveres over oftest også muliggjør «betal-for-bruk»-modeller, som gjør teknologien tilgjengelig gjennom prismodeller som senker terskelen. Brukerens interaksjon med språkmodellen gjøres typisk via et grensesnitt i en applikasjon som leveres på en skytjenesteplattform, og som fungerer som et bindeledd mellom tjenestene med brukergrensesnitt og systemene i bakkant.

KI-tjenesteleverandører, som (f.v. figur 5) Mistral, OpenAI, Anthropic og Hugging Face, leverer applikasjonsprogrammeringsgrensesnitt (API) for integrasjon av egne eller andres språkmodeller – såkalte «inferencing API». Meta sine Llama-modeller og de norske språkmodellene fra blant annet LTG/UiO, NorwAI, Nasjonalbiblioteket, m.m. er alle tilgjengelig gjennom Hugging Face. Skytjenesteplattformer, slik som (f.v.) Microsoft Azure, Google Cloud, OVH Cloud, og Amazon Web Services, leverer plattformer med ulike komponenter, for å utvikle skybaserte nettjenester, slik som digitale læringsressurser.



Figur 5 – Integrasjoners rolle i å tilgjengeliggjøre KI-teknologi i brukerrettede tjenester.

Utvexling av informasjon med plattformen og KI-teknologien foregår gjennom et applikasjonsprogrammeringsgrensesnitt (API), som er et sett med instruksjoner for å nå en adresse med ulike funksjoner som kan utføres der, og en nøkkel for å låse opp døren. Slike API gjør ny teknologi relativt enkelt å implementere, selv med begrenset kunnskap om teknologien som implementeres; dette fordi en utvikler ikke trenger å kjenne den underliggende teknologien som implementeres – kun implementasjonslaget (figur 5). Dette kan fronte en instrumentell, heller enn kritisk, tilnærming til spredning og adopsjon av teknologi. Denne vurderingen av personvernkonsekvenser utgjør slikt sett et bidrag til å ivareta et helhetlig perspektiv på implementering av en fremvoksende teknologi, slik som ulike former for kunstig intelligens.

5 PERSONOPPLYSNINGSVERN I KONTEKST AV OPPLÆRING

5.1 BEHANDLINGSGRUNNLAG

Med grunnlag i opplæringslova med forskrifter, behandles personopplysningene til elever og lærlinger for pedagogiske og skoleadministrative formål, f.eks. undervisvurdering (læringsdialog som fremmer lærelyst og vurderer vist kompetanse), dialog om sosial utvikling, og til organisering av elevens opplæring. Når en læringsressurs eller fagsystem inkluderer kunstig intelligens kan det potensielt behandles personopplysninger til å tilpasse innhold i tjenesten til elevens preferanser eller læringsarbeid, samt også lærers arbeid med undervisvurdering.

Behandling av personopplysninger krever et lovlig behandlingsgrunnlag. Det primære behandlingsgrunnlaget vil være et av alternativene i GDPR artikkel 6(1) bokstav a) til f), eller et av alternativene i GDPR artikkel 9(2) dersom det er snakk om særlig kategorier personopplysninger. Det primære behandlingsgrunnlaget som er mest aktuelt for skoleeier, er GDPR artikkel 6(1) bokstav e)⁴ eller GDPR artikkel 9(2) bokstav g).⁵ Det kreves et supplerende behandlingsgrunnlag i nasjonal lovgivning for begge disse primære behandlingsgrunnlagene.

Opplæringslova §25-1 åpner for at personopplysninger kan behandles for å utføre oppgaver når det er nødvendig for å utføre lovpålagte oppgaver. Skoleeiers supplerende behandlingsgrunnlag vil dermed bli denne bestemmelsen, jmfør den konkrete bestemmelsen i opplæringslova eller i de tilhørende forskriftene som pålegger skoleeier en oppgave, for eksempel undervisvurdering og sluttvurdering. Ettersom læreplanene er forskrifter, vil blant annet formål, nødvendighet og grunnlag for behandlinger i digitale læringsressurser kunne variere mellom de ulike fagenes læreplaner. Læreplanene for vg2 og vg3 er generelt sett mer spesifikke når det gjelder å knytte kompetansemål til oppgaver og funksjoner. Blant annet gjelder dette for læring om og med kunstig intelligens eller andre digitale teknologier. Dette har implikasjoner for formulering av nødvendighet.

5.2 PERSONVERNRISIKO

Når man vurderer personvernrisikoer i digitale læringsressurser og pedagogiske fagsystem, er det vanlig å ta med følgende risikoer:⁶

- risikoen for at personopplysninger behandles med feil,

⁴ Behandlingen av personopplysninger «er nødvendig for å utføre en oppgave i allmennhetens interesse eller utøve offentlig myndighet som den behandlingsansvarlige er pålagt.»

⁵ Behandlingen av personopplysninger «er nødvendig av hensyn til viktige allmenne interesser, på grunnlag av unionsretten eller medlemsstatenes nasjonale rett som skal stå i et rimelig forhold til det mål som søkes oppnådd, være forenlig med det grunnleggende innholdet i retten til vern av personopplysninger og sikre egnede og særlige tiltak for å verne den registrertes grunnleggende rettigheter og interesser.»

⁶ Det danske Datatilsynet har laget en liste over ti typiske brudd på personopplysningsikkerheten (Datatilsynet, 2023a), samt en veileder for kommunal sektors bruk av kunstig intelligens (Datatilsynet, 2023b). Disse generelle veilederne kan konkretiseres ytterligere ved å sette både generelle og særskilte risikoer inn i en sektoriell kontekst.

- risikoen for at det behandles for mange personopplysninger i forhold til det som er nødvendig for å oppnå formålet med behandlingen,
- risikoen for brudd på konfidensialitet, tilgjengelighet eller integritet,
- risikoen for at personopplysninger viderebehandles til nye formål uten at det er vurdert om dette er lovlig, og
- risikoen for at personopplysningene kan brukes til å fatte beslutninger som ikke er til gode for eleven.

Personopplysningsvernet i opplæringssektoren kan utfordres av blant annet:

- nødvendighet og riktighet ved bruk av elevens personopplysninger til å vurdere og å tilpasse opplæringen til eleven,
- elevens mulighet til å forstå hvilke formål skolen eller læreren bruker personopplysningene til, og
- systematisert behandling av personopplysninger for skoleadministrative oppgaver.

6 RISIKO FOR BEHANDLING AV PERSONOPPLYSNINGER VED ELEVER OG ANSATTES BRUK AV KUNSTIG INTELLIGENS I OPPLÆRINGEN

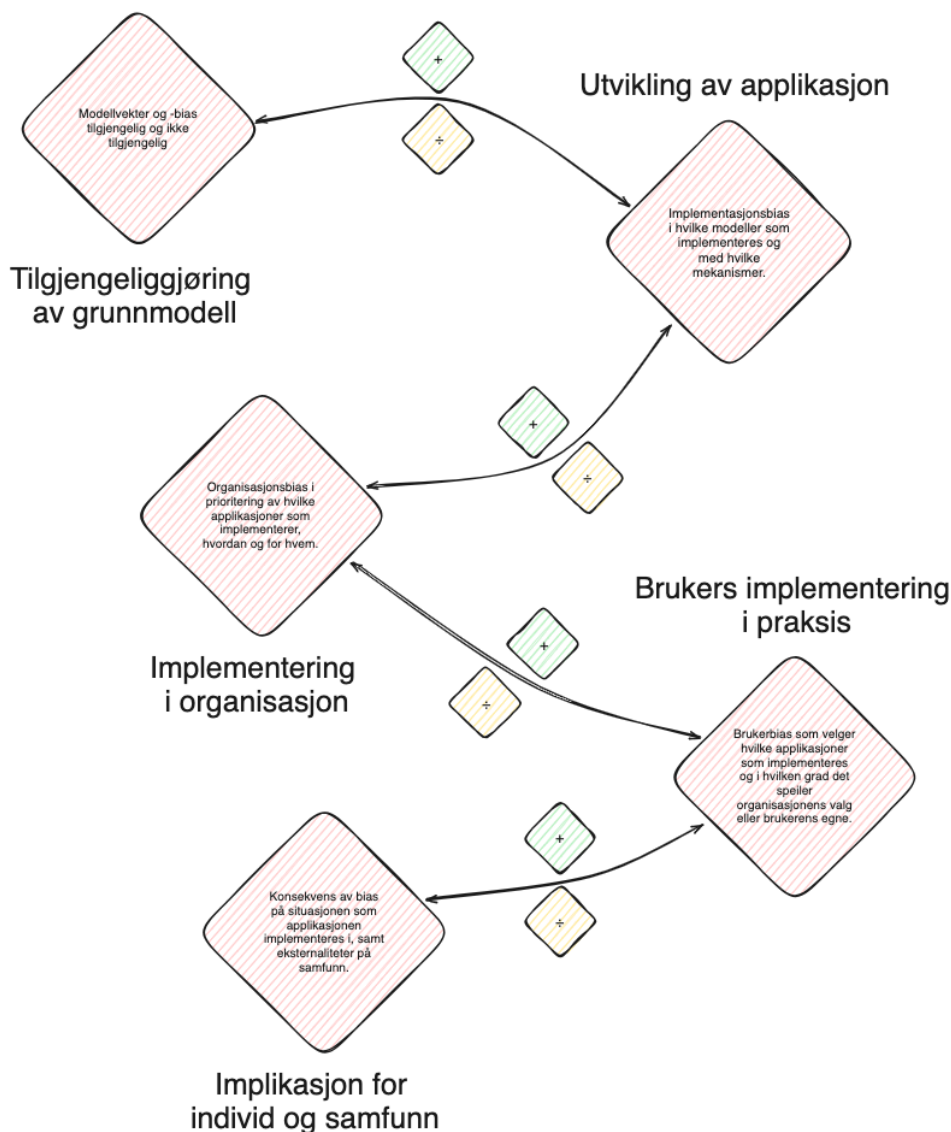
I vurdering av konsekvenser for den registrertes rett til vern av personopplysninger og øvrige rettigheter og friheter jf. artikkel 35 av personvernforordningen (GDPR), er det nødvendig å innta et helhetlig perspektiv på behandling av personopplysninger – både teknisk og organisatorisk – på ulike nivå av tjeneste og organisasjon. Det innebærer å ha en oversikt og kontroll på hvordan teknologi og praksis er forbundet, og hvordan valgene som er tatt underveis innehar ulike risikofaktorer, slik som vektning i modellen som får betydning for koblingen mellom inn- og utdata. Dette kan illustreres på ulike måter, men i denne sammenheng kan det være nyttig å følge valgene fra ytterpunkt til ytterpunkt (Figur 3) – fra utvikling av KI-modell og applikasjon, til og fra brukerens praksis og implikasjoner for den registrerte og samfunn. Sistnevnte innebærer en avveining av konsekvenser og nødvendighet av å utføre en oppgave i allmennhetens interesse og individets interesser, slik som retten til opplæring og utdanning, tanke- og ytringsfrihet, jf. GDPR artikkel 4.

Personopplysninger som har vært del av treningsdataene og dermed vært eksempler i trening, er på ulikt vis representert i modellen, og kan ikke uten videre «slettes». Så lenge de er representert i modellvektene, så kan de «trekkes ut», gjennom menneskelig eller maskinell interaksjon med modellen. Slik «memorization» (Carlini et al 2020, Thakkar et al, 2021, Kharitonov et al, 2021) diskuteres i kapittel 6.1. Det er særlig risiko dersom en modell kommer på avveie (Vincent, 2023), og deretter kan brukes i et miljø der det er mulig å forbigå sikkerhetsmekanismer som er implementert for å hindre modellen fra å gjengi opplysningene, eller der slike mekanismer er utilstrekkelige (White, 2023). Implikasjonen av dette er at så snart en modell er tilgjengelig, så er i praksis retten til sletting (artikkel 17, GDPR) eller rett til begrensning (artikkel 18, GDPR) ikke lenger mulig å oppfylle, med mindre behandlinger gjennomført ved bruk av modellen opphører i sin helhet.

Faktorer som øker eller senker risikonivå oppstår gjennom flere prosesser på ulike nivå og ved ulik tid. Disse faktorene kan f.eks. medføre:

- Manglende åpenhet om behandlingsaktiviteter
- Manglende forklarbarhet i KI-skapt innhold
- Manglende merking av KI-skapt innhold eller innvirkning på prosess og beslutning
- KI'ens definisjonsmakt og automatiserte beslutninger og profilering
- Integrasjon med eksterne datakilder eller prosesser
- Brudd på konfidensialitet, f.eks. ved reidentifisering av pseudonymisert eller anonymiserte opplysninger, utilstrekkelig tilgangsstyring eller andre sikkerhetsmekanismer.
- Produksjon av misvisende, villedende og skadelig innhold.

I kapittel 6 drøftes seks ulike risikoområder, potensielle konsekvenser for den registrerte, samt mulige tiltak for å redusere disse risikoene.



Figur 3 - Risikofaktorer som påvirker implementering av KI-tjenester og forplantning av bias i ressurs og prosess

6.1 MANGLENDE ÅPENHET I MASKINLÆRINGSMODELLER OG TJENESTER FOR KUNSTIG INTELLIGENS

Bruk av kunstig intelligens til å skape tilpasset innhold til lærere og elever skal gjøres på en måte som er forståelig og forutsigbar når dette gjøres på grunnlag av personopplysninger – slik som persontilpasset respons. Manglende informasjon og forklarbarhet rundt behandlingsaktiviteter fører til at den registrerte ikke opplever åpenhet rundt behandling av personopplysninger, og at det blir vanskelig å utøve egne rettigheter. Det kan i tillegg føre til at tilliten til både systemene,

skolen og kommunal sektor som offentlig myndighet svekkes. Manglende åpenhet er også en risiko når teknologien anvendes som beslutningsstøtte, og det ikke er mulig å forklare på hvilket grunnlag en anbefaling er generert. Slik «black-boxing» er en risiko ved maskinlæring og kunstig intelligens, og er et resultat av både manglende evne til å gjøre slike komplekse systemer forklarbare, samt manglende innsikt og kompetanse hos sluttbruker. Forretningshensyn fører også til at tjenestetilbydere ikke nødvendigvis ønsker åpenhet om hvordan sine modeller fungerer. Fenomenet fører til mange av risikoene knyttet til teknologien (Longpre et al, 2024). Samtidig kan det være ulik oppfatning rundt hva åpenhet innebærer, samt forventning og forståelsen av hvem det er forklart og forståelig for. Tilsynelatende banal funksjonalitet kan potensielt ha betydelige personvernkonsekvenser. Teknologiselskapenes dokumentasjon og veiledninger bærer i stor grad preg av å være enten markedsføring eller tiltenkt teknisk implementering. Noen aktører har begynt å publisere innhold tiltenkt å gjøre teknologier som kryptering og vekting i modeller mer åpent og forklarbart (Apple, 2024), men det er i stor grad publikasjoner fra forskningsinstitusjoner som gir best grunnlag for å gjennomføre valide vurderinger av teknologien, selv om disse sjelden omhandler de aktuelle tjenestene der behandlingene gjennomføres.

Når en språkmodell skaper innhold basert på en instruks så vil det med dagens innsiktsmetoder være uklart hvilke kriterier som er lagt til grunn for responsen (Kuhn et al, 2023). Dette er både fordi algoritmer og nevralt nettverk som disse modellene bygger på og består av, ikke lar seg forklare på samme måte som er typisk for behandling av personopplysninger i læringsressurser og fagsystem. Modellene er ikke-deterministiske, i den forstand at de er designet for å gi varierende respons, selv med samme inndata. Det er foreslått ulike metoder for å gjøre responsen fra språkmodeller forklarbar, men disse er ikke tilgjengelig i en form som er forståelig for de fleste brukere. I tillegg krever vanlige metoder for forklarbarhet en tilgang til vektingen i språkmodellen, som sluttbrukerne vanligvis ikke får. Dette kan være en særlig problemstilling ved bruk av assistenter som fungerer som læringsagenter med respons formet på bakgrunn av ulike instruksjoner og kunnskap (Russel og Norvig, 2010). En konsekvens av manglende forklarbarhet er at det ikke er mulig for eleven eller lærer, å se hvilken del av responsen som er fra treningsdata eller trening, og hvilke funksjoner som former respons. F.eks. kan en typisk respons fra en chatbot om temaer som er sensitive være preget av moralsk normative påstander som blant annet påvirkes av det som kalles moderasjon, som er mekanismer for å begrense generering av respons baserte på det modellen har lært eller som befinner seg i data modellen har tilgang til. Slike mekanismer skal blant annet hindre generering av skadelig innhold, og kan være positivt, men ikke nødvendigvis riktig i den sammenhengen. Det kan også være et resultat av andre regelsett, slik som søk i eksterne datakilder, der modellen er instruert i hva den skal svare og hvordan. Her er pedagogens rolle som fortolker viktig, men krever både verktøy og kompetanse, og vil uansett være svært begrenset av teknologiske og juridiske rammer. Den fremste faktoren er tilsynelatende manglende mulighet til å indikere hvilke deler av treningsdataene som har formet responsen til språkmodellen. Dersom modellen benyttes i et fagsystem vil det ikke være mulig med dagens verktøy og metoder, å skille autoritative kilder fra f.eks. blogger, samt i hvilken grad modellen har lært å vekte disse (Carlini, 2023 and Schaul et al, 2023). Dette kan være en negativ faktor med tanke på demokratisk medvirkning i opplæring, og at både elever og lærere skal oppleve å ha nødvendig innsikt i og kontroll over de ressursene som benyttes i opplæringen.

Dersom kunstig intelligens brukes til å utforme tilbakemeldinger til elever, så vil det være en risiko for at eleven ikke har nødvendig innsikt i om hele eller deler av en tilbakemelding er generert av kunstig intelligens. Når det er uklart for eleven hva som er lærerens vurdering av arbeidet, kan det redusere betydning av dialogen som del av undervisvurderingen.

- Det kan være uklart hvilke roller som har tilgang til logging av spørringer til språkmodeller, hos leverandøren av læringsressursen eller hos leverandøren av språkmodellen, og hva slike logger kan og blir brukt til.
- Kan bli uklart for eleven hvilken del av en periode som er vektlagt. Ulike situasjoner og bruksscenarioer vil ha ulik grad av risiko. Langsiktig systematisert bruk eller bruk i situasjoner som veier tyngre mot sluttvurdering, vil ha en annen risikoprofil enn tidvis bruk som ikke vektas av pedagogen på samme måte.
- Manglende eller for komplisert dokumentasjon, kan føre til manglende åpenhet.
- Komplekse og mange integrasjoner som ikke er implementert riktig eller fullstendig, kan medføre at tjenestene som elever og lærere tar i bruk ikke nødvendigvis fungerer som beskrevet.

En språkmodell behandler ikke treningsdata på samme måte som kilder for en søkemotor, men aktiverer vekter (representert ved distansen mellom ord) basert på inndata (tokens), gjennom en prosess som kalles «byte-pair-encoding» (Gage, 1994, Radford et al 2018). Ettersom dette er mønster av tillærte eksempler i modellens minne, heller enn hele kunnskapselementer, oppstår spørsmål knyttet til hvordan en samtalerobot kan identifisere hva som er opphav til en respons. OpenAI⁷ har tilgjengeliggjort verktøy som viser hvordan dette rent konseptuelt kan gjøres, men det vises ikke til en systematisert måte å validere og ettergå respons fra språkmodellen. Også konkurrenten Anthropic har forsøkt å gjøre responsen forklarbar gjennom å anvende en annen modell til å tolke utdata (Anthropic, 2024). Dette har implikasjoner der responsen fra språkmodellen antas å bygge på eller representere en autoritativ kilde. For eksempel der læreplanverket legges til grunn for språkmodellens respons, enten som del av treningsdata eller som søkbar ekstern datakilde, så er det uavklarte spørsmål knyttet til hvordan responsen aktiverer vekter i ulike nevroner, som igjen representeres i ulike ord i teksten den skaper. Studier indikerer at språkmodeller har en tendens til å gjengi særlig enkelte deler av treningsdataene, såkalt «memorization» (Carlini et al, 2021; Carlini et al, 2022). Dette er også en aktuell problemstilling med tanke på opphavsrett fordi språkmodellen kan reproducere nøyaktig både åndsverk og personopplysninger i treningsdata. Fenomenet betegnes som «regurgitation», eller oppgulp. Det er antydning til at språkmodeller er svært utsatt for påvirkning som kan gjøre beregning av utdata uforutsigbar utenfor et ellers kjent bruksområde (Mirzadeh et al, 2024).

Det er en vesentlig utfordring for maskinell analyse og sammenstilling, at læreplanverket er ment som et styringsinstrument som skal tolkes og anvendes av mennesker, men er tilgjengelig for trening av språkmodeller som åpne webdata. Det forutsettes slikt sett at en språkmodell nødvendigvis kan fortolke og representere læreplanverket, basert på en beregning av sannsynlighet. Læreplanverket på Utdanningsdirektoratets nettsider er samlet inn som del av treningsdataene, men er ikke nødvendigvis den eneste kilden til omtale av læreplanene på nettet, eller ikke nødvendigvis den kilden som vil være representert i de parametrene som

⁷ Selskapet som har utviklet ChatGPT

aktiveres i modllen. Blogger, nyhetsartikler og foruminnlegg kan alle tillegges uforholdsmessig stor vekt, fremfor lovtekster og regulering.

Eksempler på responser fra ulike språkmodeller:

Tabell 1 - Respons fra GPT-3.5-turbo, GPT-4o og normistral-7b-warm-instruct på instruksen "Hvordan kan undervisvurdering bidra til å fremme læring?"

Autorativ kilde	Definisjon
<p>Tekst om undervisvurdering fra udir.no</p>	<p>Undervisvurdering</p> <p>Undervisvurderingen skal bidra til å fremme læring og til å utvikle kompetanse i faget. Elevene viser og utvikler kompetanse i naturfag på vg1 studieforberedende utdanningsprogram når de bruker fagspråk, teorier og modeller til å utforske, beskrive, forklare og drøfte sammenhenger i og mellom naturfaglige fenomener. De viser og utvikler også kompetanse når de argumenterer for metodevalg, gjør etiske og sikkerhetsmessige vurderinger, reflekterer over funn og kritisk vurderer kilder og informasjon knyttet til egne utforskinger og forsøk. Videre viser og utvikler de kompetanse når de anvender og reflekterer over naturvitenskapelige praksiser og tenkemåter og naturvitenskapelig teknologi.</p> <p>Læreren skal legge til rette for elevmedvirkning og stimulere til lærelyst gjennom å legge til rette for varierte, praktiske og utforskende arbeidsmåter. Læreren og elevene skal være i dialog om elevenes utvikling i naturfag. Elevene skal få mulighet til å utforske og prøve seg fram. Med utgangspunkt i kompetansen elevene viser, skal de få mulighet til å sette ord på hva de opplever at de får til, og reflektere over egen faglig utvikling. Læreren skal gi veiledning om videre læring og tilpasse opplæringen slik at elevene kan bruke veiledningen for å utvikle kompetansen sin i naturfag.</p>
<p>Språkmodell</p>	<p>Aktiviseringer basert på instruks: «Hvordan kan undervisvurdering bidra til å fremme læring?»</p>
<p>Modell: OpenAI GPT-3.5-turbo</p> <p>Inference endepunkt: OpenAI API</p>	<p>ASSISTANT</p> <p>Undervisvurdering kan bidra til å fremme læring på flere måter:</p> <ol style="list-style-type: none"> 1. Kontinuerlig tilbakemelding: Gjennom undervisvurdering får elevene regelmessig tilbakemelding på sitt arbeid, slik at de kan identifisere styrker og områder de kan forbedre seg på. Dette bidrar til å øke bevisstheten om egen læring og fremmer en kontinuerlig forbedringsprosess. 2. Tilpasset læring: Ved å bruke undervisvurdering kan læreren tilpasse undervisningen og tilbakemeldingene til den enkelte elevs behov. Dette gjør at elevene kan få mer målrettet støtte og veiledning som hjelper dem å nå sine læringsmål. 3. Styrker motivasjonen: Ved å gi konstruktiv tilbakemelding underveis, hjelper læreren elevene til å se fremgang og suksess i sitt arbeid. Dette kan bidra til å øke motivasjonen og engas

<p>Modell: OpenAI GPT-4-turbo</p> <p>Inference endepunkt: OpenAI API</p>	<p>ASSISTANT</p> <p>Undervisvurdering, også kjent som formativ vurdering, spiller en avgjørende rolle i læringsprosessen ved å fremme læring på flere måter:</p> <ol style="list-style-type: none"> 1. **Umiddelbar Tilbakemelding:** En av de mest verdifulle aspektene ved undervisvurdering er at den gir umiddelbar tilbakemelding til både elever og lærere. Elever får innsikt i hva de gjør bra og hvor de har forbedringspotensial. Dette kan motivere dem til å jobbe videre med sine svakheter og bygge på sine styrker. 2. **Tilpasset Undervisning:** Lærere kan bruke informasjonen fra vurderingene til å justere undervisningen etter elevenes behov. Dette kan innebære alt fra å endre undervisningsmaterialet, tempoet eller gi ekstra støtte til de som trenger det. 3. **Egenvurdering og Refleksjon:** Elever blir ofte oppmuntret til å vurdere sitt eget arbeid som en del av undervisvurderingen. Dette bidrar til å utvikle elevenes evne til selvrefleksjon og selvregulering, noe som er viktig for livslang læring.
<p>Modell: NoraLLM normistral-7b-warm-instruct</p> <p>Inference endepunkt: HuggingFace inference API</p>	<p>Undervisvurdering kan være et kraftig verktøy for å fremme læring på flere måter. Først og fremst gir det lærere og elever muligheten til å vurdere en students forståelse og fremgang gjennom hele skoleåret. Dette kan hjelpe lærere med å identifisere områder der en student kan slite og gi målrettet støtte. For eksempel, hvis en student sliter med et bestemt konsept i matematikk, kan læreren gi ytterligere ressurser eller tilby ekstra hjelp for å hjelpe studenten med å forbedre seg.</p> <p>For det andre kan undervisvurdering hjelpe lærere med å tilpasse undervisningsmetodene sine for å bedre møte behovene til hver enkelt student. Ved å vurdere en students forståelse og læringsstil kan lærere lage leksjoner og aktiviteter som er mer engasjerende og effektive. Dette kan føre til økt motivasjon og engasjement hos elevene, noe som igjen kan føre til bedre læring.</p> <p>Til slutt kan undervisvurdering hjelpe lærere med å gi formativ vurdering, som er en type vurdering som er ment å hjelpe studenter med å forbedre seg. Formativ vurdering kan hjelpe lærere med å identifisere områder der en student trenger å forbedre seg og gi veiledning om hvordan de kan gjøre det. For eksempel, hvis en student sliter med grammatikk i engelsk, kan læreren gi tilbakemelding på studentens skriftlige arbeid og foreslå måter studenten kan forbedre grammatikken sin på.</p> <p>I sum kan undervisvurdering være et verdifullt verktøy for å fremme læring fordi det gir lærere og elever muligheten til å vurdere en students forståelse og fremgang, tilpasse undervisningsmetodene sine, og gi formativ vurdering for å hjelpe studenter med å forbedre seg. Ved å bruke undervisvurdering på en effektiv måte, kan lærere skape et mer engasjerende og effektivt læringsmiljø som fører til bedre læring for alle studenter.</p>

Det er tilsynelatende lav variasjon i ordlyden i de ulike responsene, men det er verdt å merke seg at begreper som oftere er assosiert med organisasjonslæring, også blir skapt i responsen (fig.6.1). Ord som «suksess» og begrep som «kontinuerlig forbedring» gir et noe annet bilde av undervisvurdering enn det som ligger til grunn i læreplanverket. Språkmodellene er ofte heller ikke konsistente med begreper, slik som «elev» og «student», eller ulike perspektiv på begrepet «læring». Dette er i språkmodellen en faktor av ordenes vektning (avstand til hverandre), og dermed også hva modellen vil skape av respons. Dette illustrerer behovet for å utforme instruksjoner slik at språkmodellene vektet hensiktsmessig, men er vanskelig uten å kjenne modellens treningsdata, vekter og bias – samt kunne tolke disse på en hensiktsmessig måte.

Betydningen av hva en språkmodell er trent på og hva den er trent til å vekte i parameterne illustreres ved særnorske ord og uttrykk der små nyanser kan utgjøre vesentlig forskjell i ordenes betydning og kontekst. I figur 6.1.2 blir språkmodellene instruert til: «Lag en kort disposisjon på hvordan et stev bør se ut». OpenAI sine språkmodeller tolker input-tokenet «stev» å høyst sannsynlig være forbundet til enten «stevne» eller «stevning». Nora-modellen tolker korrekt, at input-tokenet «stev» er en vise, dikt eller rim. I figur 6.1.3 illustreres lignende funksjon

ved instruksen «Hva vil det si å røkte en fisk». OpenAI sine språkmodeller tolker input-tokenet «røkte» til å mest sannsynlig være forbundet med å tilberede fisk. Nora-modellen tolker korrekt, at det er oppdrett og stell av fisk det er snakk om.

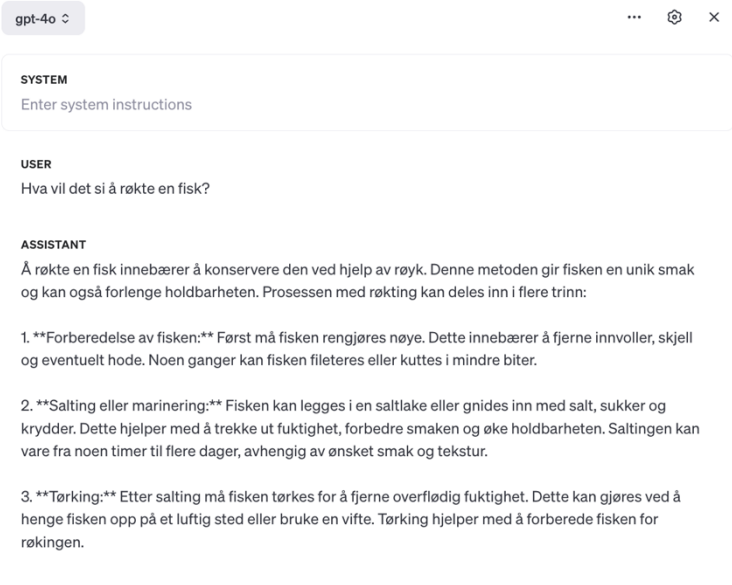
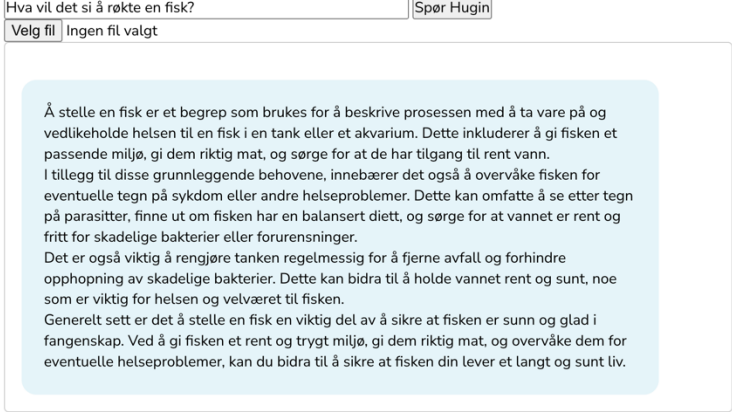
Tabell 2 - Respons fra GPT-3.5-turbo, GPT-4o og normistral-7b-warm-instruct på instruksen "Lag en kort disposisjon på hvordan et stev bør se ut"

Språkmodell og endepunkt for interaksjon	Aktiveringer basert på instruks «Lag en kort disposisjon på hvordan et stev bør se ut»
<p>Modell: OpenAI GPT-3.5-turbo Inference endepunkt: OpenAI API</p> <p>Tokenet «stev» aktiverer nært «stevne».</p>	<div style="border: 1px solid #ccc; padding: 10px;"> <p>USER Lag en kort disposisjon på hvordan et stev bør se ut.</p> <div style="background-color: #f0f0f0; padding: 10px; border: 1px solid #ccc;"> <p>ASSISTANT</p> <ol style="list-style-type: none"> 1. Innledning: <ul style="list-style-type: none"> - Presenter stevets tema og formål - Velkommen til deltakere og publikum - Introduksjon av stevets dommere og program 2. Konkurransedelen: <ul style="list-style-type: none"> - Start med de ulike konkurransekategoriene og deltakerne - Konkurrentene opptre og viser frem sine ferdigheter - Dommerne gir tilbakemelding og scorer deltakerne 3. Underholdningsdelen: <ul style="list-style-type: none"> - Innslag av gjesteartister, musikalske opptredener eller andre underholdningsinnslag - Skape en morsom og engasjerende stemning 4. Avslutning: <ul style="list-style-type: none"> - Kåring av vinnere i de ulike kategoriene - Takk til deltakere, dommere og frivillige - Oppmuntre til videre deltakelse og interesse for stevets formål 5. Sosialt samvær: <ul style="list-style-type: none"> - Mulighet for deltakere og publikum til å mingle og utveksle erfaringer - Mat og drikke serveres for å skape en </div> </div>

<p>Modell: OpenAI GPT-4-turbo Inference endepunkt: OpenAI API</p> <p>Tokenet «stev» aktiverer nært «stevning».</p>	<div style="border: 1px solid #ccc; padding: 10px;"> <p>USER Lag en kort disposisjon på hvordan et stev bør se ut.</p> <hr/> <p>ASSISTANT Enter an assistant message here.</p> <hr/> <p>En disposisjon for et stev kan struktureres på følgende måte for å sikre at innholdet er klart, koherent og overbevisende. Stevet kan her tenkes som et formelt dokument eller innkalling benyttet i juridisk sammenheng, så som en retts sak:</p> <ol style="list-style-type: none"> Tittel: Gi stevet en passende tittel som reflekterer innholdet. Eksempel: "Stevning i Saken [Navn] mot [Navn]" Innledning: <ul style="list-style-type: none"> Formål: Kort forklaring av hva dokumentet er, og hva det søker å oppnå eller avklare. Referanse til relevant lovverk: Angi hvilken lov eller juridisk ramme som gjelder for saken. Parter: <ul style="list-style-type: none"> Saksøker: Navn og kontaktinformasjon. Saksøkte: Navn og kontaktinformasjon. Faktisk bakgrunn: <ul style="list-style-type: none"> Kort beskrivelse av de faktiske omstendighetene som har ført </div>
<p>Modell: NoraLLM normistral-7b-warm-instruct Inference endepunkt: Ollama/MacOS</p> <p>Tokenet «stev» aktiverer nært vise og dikt.</p>	<div style="background-color: #1a2b3c; color: white; padding: 10px;"> <p>>>> Lag en kort disposisjon på hvordan et stev bør se ut. Et stev er en type firelinjet folkeviser som ofte brukes til å uttrykke en bestemt stemning eller følelse. Her er en kort disposisjon av hva et stev vanligvis inneholder:</p> <ol style="list-style-type: none"> Tittel: Tittelen er den første linja i stevet, og den setter tonen for resten av visa. Den kan være humoristisk, seriøs, eller til og med litt rampete. Omkved: Omkvedet er den andre linja i stevet, og det gjentas etter hver eneste linje. Omkvedet er ofte det mest minneverdige eller fengende i stevet, og det hjelper til med å skape en fengende melodi. Kor: Koret er den tredje linja i stevet, og det er der hovedhandlingen eller temaet i stevet finner sted. Koret gjentar ofte et bestemt ord eller frase, og det brukes til å forsterke temaet i stevet. Refreng: Refrenget er den fjerde linja i stevet, og det gjentas etter hver eneste linje. Refrenget er ofte det mest catchy eller minneverdige i stevet, og det hjelper til med å skape en fengende melodi. <p>Jeg håper dette hjelper! Si ifra hvis du har noen andre spørsmål om stev eller andre typer folkeviser.</p> </div>

Tabell 3 - Respons fra GPT-3.5-turbo, GPT-4o og normistral-7b-warm-instruct på instruksjonen "Hva vil det si å røkte en fisk?"

Språkmodell	Aktiveringer basert på instruks «Hva vil det si å røkte en fisk?»
<p>Modell: OpenAI GPT-3.5-turbo Inference endepunkt: OpenAI API</p> <p>Aktiverer nært røking og tilberedelse av mat.</p>	<div style="border: 1px solid #ccc; padding: 10px;"> <p style="font-size: small;">gpt-3.5-turbo ▾</p> <hr/> <p>SYSTEM Enter system instructions</p> <hr/> <p>USER Hva vil det si å røkte en fisk?</p> <hr/> <p>ASSISTANT Å røke en fisk refererer til en matlagingsteknikk der fisken blir utsatt for røyk fra røykbriketter eller flis for å tilføre en røkt smak. Prosessen innebærer vanligvis å salte fisken før den blir hengt opp i en røykeovn eller røykekasse, der den blir utsatt for lav varme og røyk i flere timer. Dette gir fisken en karakteristisk smak og aroma som mange finner tiltalende. Røkt fisk kan spises som den er, brukes i ulike retter eller snacks. Det er en vanlig matlagingsspraksis over hele verden og er spesielt populært i kystområder hvor fisk er en viktig del av kostholdet.</p> </div>

<p>Modell: OpenAI GPT-4o Inference endepunkt: OpenAI API</p> <p>Aktiverer nært røking og tilberedelse av mat.</p>	 <p>The screenshot shows a chat window with the following content:</p> <p>SYSTEM Enter system instructions</p> <p>USER Hva vil det si å røkte en fisk?</p> <p>ASSISTANT Å røkte en fisk innebærer å konservere den ved hjelp av røyk. Denne metoden gir fisken en unik smak og kan også forlenge holdbarheten. Prosessen med røking kan deles inn i flere trinn:</p> <ol style="list-style-type: none"> 1. Forberedelse av fisken: Først må fisken rengjøres nøye. Dette innebærer å fjerne innvoller, skjell og eventuelt hode. Noen ganger kan fisken fileteres eller kuttes i mindre biter. 2. Salting eller marinering: Fisken kan legges i en saltlake eller gnides inn med salt, sukker og krydder. Dette hjelper med å trekke ut fuktighet, forbedre smaken og øke holdbarheten. Saltingen kan vare fra noen timer til flere dager, avhengig av ønsket smak og tekstur. 3. Tørking: Etter salting må fisken tørkes for å fjerne overflødig fuktighet. Dette kan gjøres ved å henge fisken opp på et luftigt sted eller bruke en vifte. Tørking hjelper med å forberede fisken for røkingen.
<p>Modell: NoraLLM normistral-7b-warm-instruct Inference endepunkt: HuggingFace inference API</p> <p>Aktiverer nært fiskeoppdrett og pass av fisk.</p>	 <p>The screenshot shows a chat window with the following content:</p> <p>Hva vil det si å røkte en fisk? Spør Hugin</p> <p>Velg fil Ingen fil valgt</p> <p>Å stelle en fisk er et begrep som brukes for å beskrive prosessen med å ta vare på og vedlikeholde helsen til en fisk i en tank eller et akvarium. Dette inkluderer å gi fisken et passende miljø, gi dem riktig mat, og sørge for at de har tilgang til rent vann. I tillegg til disse grunnleggende behovene, innebærer det også å overvåke fisken for eventuelle tegn på sykdom eller andre helseproblemer. Dette kan omfatte å se etter tegn på parasitter, finne ut om fisken har en balansert diett, og sørge for at vannet er rent og fritt for skadelige bakterier eller forurensninger. Det er også viktig å rengjøre tanken regelmessig for å fjerne avfall og forhindre opphopning av skadelige bakterier. Dette kan bidra til å holde vannet rent og sunt, noe som er viktig for helsen og velværet til fisken. Generelt sett er det å stelle en fisk en viktig del av å sikre at fisken er sunn og glad i fangenskap. Ved å gi fisken et rent og trygt miljø, gi dem riktig mat, og overvåke dem for eventuelle helseproblemer, kan du bidra til å sikre at fisken din lever et langt og sunt liv.</p>

Eksemplene illustrerer viktigheten av forklarbarhet for å øke graden av åpenhet. Der responsen fra OpenAI sine modeller ikke kan forklares ved å tolke modellenes vektning og bias, så er både modellvektene og treningsdataene til Nora-modellene åpent tilgjengelig. Dette understreker et behov for både en åpen og forklarbar infrastruktur, samt undersøkelsesmetoder som kan bidra til dette. Bruk av en type maskinlæringsmodeller som kalles autoencodere til å tolke transformer-modeller (Anthropic, 2024. Gao et al, 2024) har vist seg å ha potensiale som metoder for forklarbarhet, men krever åpne modeller og resulterer ikke i konsepter som kan fortolkes av en typisk sluttbruker av KI-tjenester. Dette, en åpen infrastruktur for interaksjon med språkmodeller (inferens), åpne språkmodeller og valide metoder for forklarbarhet, er nødvendig for å kunne ivareta åpenhetsprinsippet i personvernforordningen (GDPR).

6.1.1 Risiko og konsekvens ved manglende åpenhet

ID	Risiko	Konsekvens	Mulig tiltak
6.1.01	Manglende mulighet til å forklare responsen/utdata fra kunstig intelligens.	Den registrerte får ikke god nok informasjon om behandlinger med kunstig intelligens til å kunne ivareta egne rettigheter. Den som er ansvarlig for behandlingen har ikke god nok informasjon til å ivareta den registrertes rettigheter. Eksempelvis så kan manglende forklarbarhet bidra til at KI tas i bruk på måter som svekker retten til opplæring og utdanning, og videre arbeid.	Implementere kunstig intelligens på måter som gir økt grad av åpenhet og forklarbarhet, og velge teknologi som i større grad muliggjør dette.
6.1.01	Utilstrekkelig informasjon gjør at elev og lærer har liten mulighet til å få innsikt i behandlingene, og å teknologien som brukes og hvordan.	Den registrerte er ikke kjent med negative faktorer ved teknologi og praksis som kan få betydning for tanke- og ytringsfrihet. For eksempel ved bruk av KI-systemer som skaper feilaktig eller misvisende respons.	God dokumentasjon av formål og midler ved behandlingen, og at informasjon til de registrerte er tilstrekkelig forståelig.
6.1.01	Utilstrekkelig merking av KI-generert innhold som presenteres for en elev i eller utenfor en læringsressurs kan føre til manglende innsikt i valg som påvirker egen læring.	Den registrerte oppfatter ikke at den er eller har vært gjenstand for en behandling med kunstig intelligens, noe som påvirker mulighetene til å forstå og avdekke feil eller mangler, og som påvirker egne valg.	Tydlig merking i brukergrensesnitt og i informasjon til de registrerte, der den er gjenstand for en behandling som utføres med KI.

6.2 FRA BESLUTNINGSSTØTTE TIL BESLUTNING

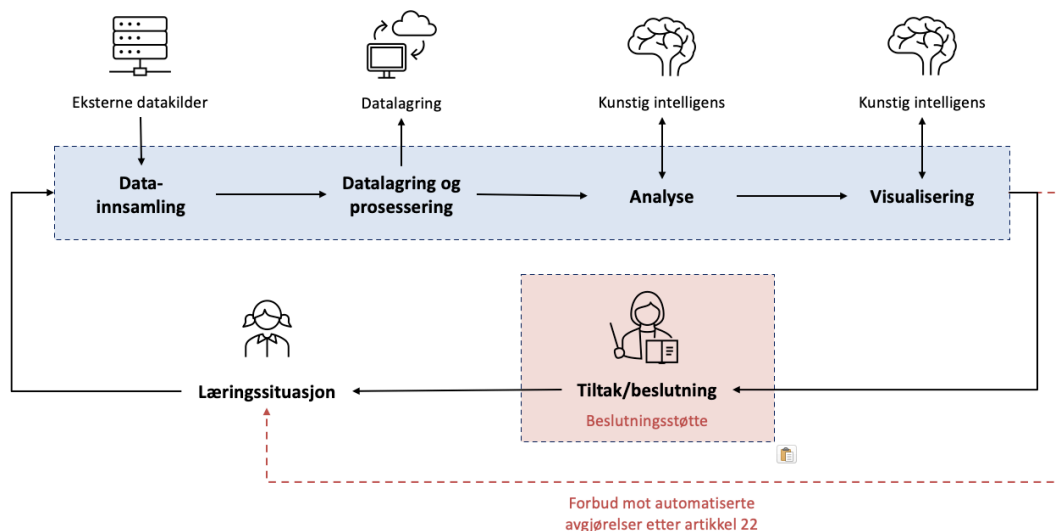
Fagsystemer og digitale læringsressurser vil i mange tilfeller ha funksjonalitet som kan støtte lærer og elev på ulike måter. Disse kan bygge på både helt eller delvis automatiserte behandlinger av personopplysninger. Beslutningsstøtte i opplæring kan forstås som systemer for læringsanalyse, som «benytter digital teknologi til å sortere, analysere og tolke data for å finne ny kunnskap og få nye innsikter» (NOU2023:19, s. 9), da rettet mot elevens læring. Både elev og lærer kan nyttiggjøre læringsanalyse, gjennom f.eks. adaptive læremidler eller andre typer anbefalingssystemer, i læring og underveisvurdering⁸, og legge systemets vurderinger mer eller mindre til grunn for videre beslutninger.

Digitale læringsressurser gjør det mulig for lærere å periodisere oppgaver og gi summative tilbakemeldinger. Der disse ressursene brukes oftere og spenner lengre perioder, kan de gi uforholdsmessig stor vektning til automatisk scoring (typisk i prosent eller annen tallverdi), som omsettes til karakterer. Dette kan ha betydelig innvirkning på underveisvurdering og

⁸ Jf. forskrift til opplæringslova § 9-11.

standpunkt karakter, uten at læreren nødvendigvis har innsyn i faktorene som er lagt til grunn. Dermed blir det vanskelig for læreren å tolke og korrigere anbefalingene. Samtidig er det heller ikke sikkert eleven har innsyn i og forståelse av hvordan systemet har vurdert og vektet eller at systemet har blitt brukt i undervisningsvurdering (figur 1).

6.2.1 Artikkel 22: Automatiserte beslutninger og profilering



Tilpasset fra: [NOU 2023:19 Læring, hvor ble det av deg i alt mylderet](#)



Figur 4 - Konseptualisering av læringsanalyse i NOU 2023:19

Personvernforordningen artikkel 22 forbyr avgjørelser som utelukkende er basert på automatisert behandling, inkludert profilering, og som får rettslig betydning eller i betydelig grad påvirker den avgjørelsen gjelder. En avgjørelse er utelukkende basert på automatisert behandling der det ikke foreligger noen meningsfull menneskelig involvering⁹, og forbudet i artikkel 22 gjeldende der behandlingen får rettsvirkning eller i betydelig grad påvirker eleven (Fig. 6.2.2).

Det er læreren som har ansvaret for å gi eleven undervisningsvurdering og for å fastsette standpunkt karakter i faget. Det er også læreren som er kompetent til å endre beslutningen, men det er behandlingsansvarlig som skal påse at det gjøres og at den registrertes rettigheter ivaretas.

⁹ Se Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (wp251rev.01) s. 21. Meningsfull menneskelig involvering foreligger der en person med nødvendig autoritet og kompetanse har tilgang på alle relevante data for å endre beslutningen – og faktisk utøver denne. Behandlingsansvarlig skal identifisere grad av menneskelig involvering og i hvilken fase det finner sted.

Det er samtidig kun i tilfellene der en automatisert behandling har rettslig betydning eller i betydelig grad påvirker eleven, og der læreren ikke har noen meningsfull involvering i å endre beslutningen at artikkel 22 er gjeldende. Standpunktvurdering, som føres på vitnemål, får helt klart rettsvirkning for elev, men for underveisvurdering er det noe vanskeligere å definere hva som i betydelig grad påvirker eleven. Dersom lærer ikke har nødvendig innsikt i eller forstår hvordan et digitalt læremiddel har vurdert en elevs prestasjon og legger denne vurderingen ukritisk til grunn for underveisvurdering eller standpunkt karakter, som i betydelig grad påvirker eleven, så kan det foreligge et brudd på personvernforordningen art. 22.

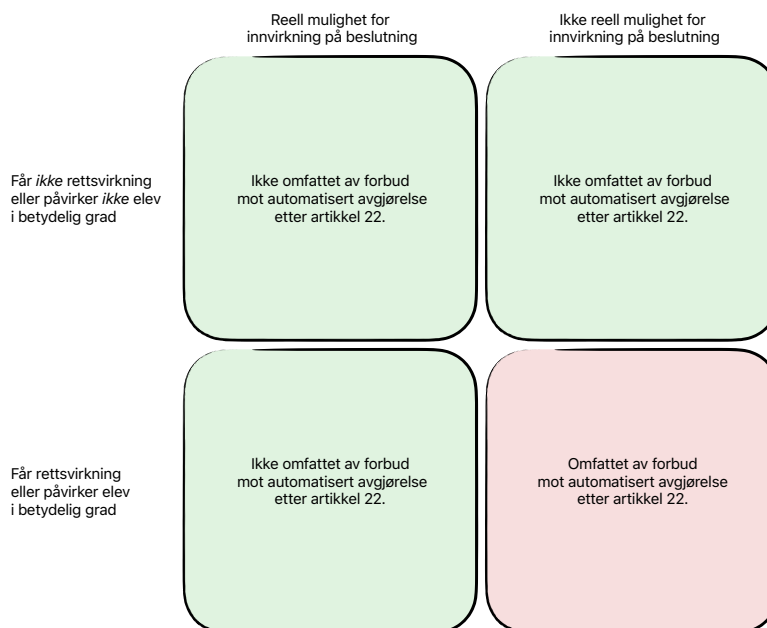
På den ene siden vil sporadisk og lav grad av KI-genererte tilbakemeldinger nok ikke ha betydelig eller varig påvirkning på elevens muligheter til å lære og til å øke sin kompetanse. Samtidig som høy grad av KI-tilbakemelding og hyppig bruk kan ha større påvirkning. Det er dog tilfeller, slik som med KI-basert kartlegging av f.eks. lese- og skrivevansker ikke nødvendigvis hyppighet eller varighet som er av betydning, men effektene det kan ha for videre muligheter for læring, og retten til opplæring og utdanning. Uavhengig av om behandlingen faller innunder forbudet i artikkel 22, så vil for eksempel bruk av kunstig intelligens til å profilere elevens arbeidsprestasjon være en behandling med høy risiko. Åpenhet i KI-tjenesten og muligheten til å forklare usikkerhet i semantisk variasjon er derfor viktig for å tolke genererte tilbakemeldinger (Kuhn et al, 2023).

Eleven benytter et utvalg av ressurser og metoder i sin læring og læreren vil vurdere eleven på flere og varierte måter. Dette grunnprinsippet for underveisvurdering vil kunne gi læreren både den nødvendige autoritet og gi rett på nødvendig krav til vurderingskompetanse, for å kunne gi underveisvurdering med en bruk av kunstig intelligens som ikke omfattes av forbudet. Det største hinderet er kravet til å vurdere alle relevante data, som er en gjennomgående åpenhetsutfordring ved bruk av kunstig intelligens.

“The controller cannot avoid the Article 22 provisions by fabricating human involvement. For example, if someone routinely applies automatically generated profiles to individuals without any actual influence on the result, this would still be a decision based solely on automated processing.

To qualify as human involvement, the controller must ensure that **any oversight of the decision is meaningful**, rather than just a token gesture. It should be **carried out by someone who has the authority and competence to change the decision**. As part of the analysis, they should **consider all the relevant data**.

As part of their DPIA, the controller should **identify and record the degree of any human involvement** in the decision-making process and **at what stage this takes place**.” (WP251, 2018, s.21)



Figur 5 - Forslag til illustrasjon av kriterier for artikkel 22

Tabell 4 - Eksempler der bestemmelsen om forbud mot automatiserte avgjørelser kan få innvirkning

ID	Behandlingsaktivitet	Ansvarliges autoritet	Ansvarliges kompetansebehov	Ansvarliges mulighet for innsikt i relevante data
6.2.1.01	Lese, skrive og lyttekartlegging basert på maskinell analyse, som danner grunnlag for enkeltvedtak eller for vesentlige valg tatt i opplæringen.	Den som gjennomfører en kartlegging eller skal nyttiggjøre seg av resultatene, skal ha muligheten til å overprøve eventuell vurdering eller scoring, helt eller delvis.	Behov for tilstrekkelig feltspesifikk kompetanse (f.eks. pedagogisk, didaktisk språkfaglig kompetanse) Kompetanse om kunstig intelligens i de pedagogiske fagsystemene de benytter.	Behov for innsikt i data som viser fonetiske, morfologiske, syntaktiske og semantiske uttrykk som fremkommer i elevens arbeid og som læreren kan se i sammenheng med en maskinell analyse. F.eks. visuelle markeringer i en rapport.
6.2.1.05	Automatisert tilbakemelding gjennom profilering av arbeidsprestasjon/elevarbeid og generering av formativ vurdering basert på analyserte data.	Lærer har det faglige ansvaret for undervisningsvurdering. Det inkluderer å definere sentrale deler av undervisningsmetodene, inkludert hvordan elevens kompetanse vurderes, jf.	Behov for tilstrekkelig profesjonsfaglig digital kompetanse, inkl. i konsekvenser ved behandling av personopplysninger i praksis, jf.	Behov for innsikt i relevante data som gir informasjon om kompetanse, læring og lærelyst.

		opplæringslova § 17-6 (Faglig ansvar for opplæringa).	§ 17-4.Krav om relevant kompetanse i undervisningsfag, samt opplæringsforskrifta § 20-2.Styrt tilgang til personopplysningar.	
6.2.1.06	Automatisert beregning av summativ vurdering gjennom profilering av arbeidsprestasjon/elevarbeid og generering av en verdi som danner grunnlag for enkeltvedtak om vurdering av sluttkompetanse.	Lærer har det faglige ansvaret for sluttvurdering, som inkluderer å vurdere elevens kompetanse i fag ved avslutning av opplæringa, jf. opplæringslova § 17-6 (Faglig ansvar for opplæringa).	Behov for tilstrekkelig profesjonsfaglig digital kompetanse, inkl. i konsekvenser ved behandling av personopplysninger i praksis, jf. § 17-4.Krav om relevant kompetanse i undervisningsfag, samt opplæringsforskrifta § 20-2.Styrt tilgang til personopplysningar.	Behov for innsikt i relevante data som gir informasjon om sluttkompetanse.

6.2.2 KI-forordningen: Høyrisikoformål og kompetansekrav

EUs Artificial Intelligence Act 2024/1689 (AIA) regulerer bruk av KI til høyrisikoformål i opplæring- og utdanning. Annex III til artikkel 6-2 kategoriserer bruk av KI i opplæring og utdanning som KI-systemer med høy risiko i bestemte tilfeller. De tilfellene er der et formål er definert i artikkel 6-1 eller 6-2, eller får betydningsfull innvirkning på en beslutning, eller profilerer (jf. artikkel 6-3). Definisjonen av profilering i KI-forordningen henviser til definisjonen i personvernforordningen (GDPR).

«profilering» enhver form for automatisert behandling av personopplysninger som innebærer å bruke personopplysninger for å vurdere visse personlige aspekter knyttet til en fysisk person, særlig for å analysere eller forutsi aspekter som gjelder nevnte fysiske persons arbeidsprestasjoner, økonomiske situasjon, helse, personlige preferanser, interesser, pålitelighet, atferd, plassering eller bevegelser,» (Artikkel 4-4, GDPR)

Vedtatt EU-lovgivning om regulering av kunstig intelligens forordning 2024/1689 (European Parliament, 2024), fortalepunkt 56:

«The deployment of AI systems in education is important to promote high-quality digital education and training and to allow all learners and teachers to acquire and share the necessary digital skills and competences, including media literacy, and critical thinking, to take an active part in the economy, society, and in democratic processes. However, AI systems used in education or vocational training, in particular for determining access or admission, for assigning persons to educational and vocational training institutions or programmes at all levels, for evaluating learning outcomes of persons, for assessing the appropriate level of education for an

individual and materially influencing the level of education and training that individuals will receive or will be able to access or for monitoring and detecting prohibited behaviour of students during tests should be classified as high-risk AI systems, since they may determine the educational and professional course of a person's life and therefore affect that person's ability to secure a livelihood. When improperly designed and used, such systems may be particularly intrusive and may violate the right to education and training as well as the right not to be discriminated against and perpetuate historical patterns of discrimination, for example against women, certain age groups, persons with disabilities, or persons of certain racial or ethnic origins or sexual orientation.» (European Parliament, 2024).

ANNEX III	
HIGH-RISK AI SYSTEMS REFERRED TO IN ARTICLE 6(2)	
High-risk AI systems pursuant to Article 6(2) are the AI systems listed in any of the following areas:	
3. Education and vocational training:	
(a)	AI systems intended to be used to determine access or admission or to assign natural persons to educational and vocational training institutions at all levels ;
(b)	AI systems intended to be used to evaluate learning outcomes, including when those outcomes are used to steer the learning process of natural persons in educational and vocational training institutions at all levels ;
(c)	AI systems intended to be used for the purpose of assessing the appropriate level of education that an individual will receive or will be able to access, in the context of or within educational and vocational training institutions ;
(d)	AI systems intended to be used for monitoring and detecting prohibited behaviour of students during tests in the context of or within educational and vocational training institutions.

KI-forordningens artikkel 4 stiller også krav til tilstrekkelig kompetanse hos brukerne av KI-systemer:

“Providers and deployers of AI systems shall take measures to ensure, to their best extent, a sufficient level of AI literacy of their staff and other persons dealing with the operation and use of AI systems on their behalf, taking into account their technical knowledge, experience, education and training and the context the AI systems are to be used in, and considering the persons or groups of persons on whom the AI systems are to be used.” (European Parliament, 2024).

AI Literacy-begrepet i lovforslaget må forstås slik at det er snakk om kompetanse, og at tradisjonen vi har for bruk av kompetansebegrepet tilsier at vi skal ha fokus på å utvikle kunnskap, ferdigheter og holdninger. I kapittel 8.9.1.2 drøftes det måter rammeverket for profesjonsfaglig digital kompetanse kan være et hensiktsmessig utgangspunkt for å ivareta kravene etter artikkel 4 i KI-forordningen.

Leverandører av KI-modeller for generelle formål er underlagt krav om å tydeliggjøre hvilke formål teknologien deres kan brukes til, samt avgrense eventuelle begrensninger i bruken. Leverandører som ikke gjør dette, risikerer selv å bli ansett som tilbydere av KI-systemer, der teknologien deres kun utgjør en komponent i en større sammenheng. Leverandører har begynt å tilpasse sine tjenestevilkår for å ivareta ansvaret deres etter KI-forordningen. Eksempelvis Microsofts «codes of conduct», for bruk av Azure AI Services:

“For chatbots that (i) are erotic, romantic, or used for erotic or romantic purposes, or which are otherwise prohibited by this Code of Conduct; (ii) are personas of specific people without their explicit consent; (iii) **claim to have special wisdom/insight/knowledge, unless very clearly labeled as being for entertainment purposes only;** or (iv) **enable end users to create their own chatbots without oversight;**”

Bruk av KI-tjenester, slik som Microsoft Azure AI Services, til å utforme chatboter som tilsynelatende skal ha innsikt i eller kunnskap om et felt, kan være i brudd med tjenestevilkår, men vil i tillegg ha implikasjoner for tilbyderen av systemet, etter blant annet åpenhets- og kompetansekravene i KI-forordningen.

Tabell 5 – Høyriskoformål i AIA artikkel 6-2, annex III, for opplæring og utdanning.

ID	Høyriskoformål	Områder for videre utforskning jf. høyrisikokrav
6.2.2.1	Inntak	KI i inntakssystemer, samt skoleeier og skolers prosesser for å oppfylle formål om inntak til videregående opplæring.
6.2.2.2	Undervisvurdering	Digitale læringsressurser og pedagogiske fagsystem med KI, samt pedagogisk praksis for undervisvurdering, som har formål om å evaluere og fremme kompetanse, læring og læringsutbytte, samt lærelyst.
6.2.2.3	Karriere- og studieveiledning	KI-systemer brukt for å profilere og utføre prediksjon av mulige eller anbefalte opplærings-, utdannings- og karrierevalg.
6.2.2.4	Prøver og eksamen	Prøve- og eksamensverktøy, slik som overvåkning i situasjonen, samt profilering og prediksjon utført i etterkant, brukt til å avdekke juks eller forsøk på juks.

6.2.3 Risiko og konsekvens ved kunstig intelligens som beslutningsstøtte

Tabell 6 - Risiko og konsekvens ved kunstig intelligens som beslutningsstøtte

ID	Risiko	Konsekvens	Mulig tiltak
6.2.2.01	Beregning fra kunstig intelligens legges til grunn for en beslutning med rettsvirkning eller som i betydelig grad påvirker eleven.	Potensielt omfattet av forbud om automatisert avgjørelse og profilering.	Kartlegge hvilke prosesser som har høy risiko for å resultere i at beregninger fra kunstig intelligens legges til grunn. Revidering av rutiner og styrende dokument. Utvikle eller anskaffe opplæringsressurser og gjennomføre opplæring.

6.2.2.02	Manglende innsikt i relevante data som legges til grunn for beregningen.	Den ansvarlige personen kan potensielt fatte en beslutning eller baserer videre beslutninger på feilaktig grunnlag.	Bedre oversikt over hvilke opplysninger som utgjør inndataene, hvilke vektorer og bias modellen har, samt automatiserte prosesser/mekanismer utenfor modellen som kan påvirke utdata. (f.eks. moderasjon)
6.2.2.03	Manglende kompetanse til å korrigere beregning gjort av kunstig intelligens.	Den ansvarlige personen kan potensielt fatte en beslutning eller baserer videre beslutninger på feilaktig grunnlag eller en feiltolkning av grunnlaget.	Kartlegge kompetanseområder der det er manglende kompetanse og iverksette opplæring individuelt eller i grupper, avhengig av behov. Utvikle eller anskaffe opplæringsressurser og gjennomføre opplæring.
6.2.2.04	Manglende autoritet til å korrigere beregning gjort av kunstig intelligens.	Den ansvarlige personen kan potensielt miste muligheten til å korrigere systemet, slik at resultatet blir et annet enn hva det er grunnlag for. F.eks. en automatisert avgjørelse.	Kartlegge styrings- og beslutningsprosess. Revidere virksamhetsrutiner og styrende dokumenter som regulerer prosessansvaret. Videreutvikle brukergrensesnitt for å gi den ansvarlige personen bedre kontroll over systemet.

6.3 VIDEREBEHANDLING UFORENLIG MED OPPRINNELIG FORMÅL OG ROLLEIMPLIKASJONER

Personopplysninger skal kun samles inn for spesifikke, uttrykkelig angitte og berettigede formål og ikke viderebehandles på en måte som er uforenlig med disse formålene, jf. personvernforordningen art. 5 nr. 1 bokstav b. Dersom et digitalt læremiddel benytter personopplysninger til å vurdere elevens læring og samtidig benytter personopplysninger til maskinlæring vil dette være to ulike formål.

En av utfordringene med kunstig intelligens er behovet for å kontinuerlig videreutvikle både modellene som ligger til grunn og tjenestene de inngår i. Dette innebærer i mange tilfeller en behandling av personopplysninger. En kjent problemstilling er at databehandler deklarerer at den kun benytter «aggregerte» eller «anonymiserte» opplysninger, uten å definere nærmere eller å beskrive metode for å aggregere eller anonymisere. Etersom det ikke nødvendigvis er behandlingsansvarlig som tar initiativ til denne videreutviklingen eller «eier» modellen eller tjenesten som videreutvikles, så risikerer databehandler å bryte med instruks fra behandlingsansvarlig, og selv bli behandlingsansvarlig for behandlingene, jf. GDPR artikkel 28-10. Det setter også skoleeier som behandlingsansvarlig i den situasjonen at det reiser tvil om det er gjort nødvendig vurderinger av databehandler jf. GDPR artikkel 28-1, og om

behandlingsansvarlig reelt sett definerer roller, formål og eventuelt grunnlag for viderebehandling.

En del leverandører av KI-som-tjeneste behandler også all inn- og utdata i en 30 dagers periode, gjennom lagring og potensiell automatisert eller manuell gjennomgang av data som har gått gjennom endepunktet, for å avdekke ulike brudd på privatrettslige vilkår eller nasjonalt lovverk. Dette kalles ofte «abuse monitoring», men kan ha formål utover å avdekke misbruk av tjenestene etter lov og forskrift, behandlingsansvarliges instruks eller lignende. Et slikt system kan ha til hensikt å avdekke brudd på privatrettslige tjenestevilkår, definert av databehandler. Skoleeier som behandlingsansvarlig har som regel ikke anledning til å instruere databehandler i om eller hvordan en slik behandling skal forekomme, og det er ikke nødvendigvis mulig å konkret vurdere vesentlighet ved slike formål og midler, grunnet manglende tilgang på informasjon.

I de fleste tilfeller der personopplysninger behandles for å gi tilgang til en digital læringsressurs med kunstig intelligens eller der den samme tjenesten skal gi tilpasset respons basert på personopplysninger til lærer eller elev, vil skoleeier være behandlingsansvarlig, og leverandør være databehandler. Databehandler skal behandle personopplysninger på instruks fra behandlingsansvarlig, og kan kun viderebehandle personopplysninger til formål som er forenelige med behandlingsansvarliges formål, jf. personvernforordningen art. 5 nr. 1 bokstav b og art. 6 nr. 4. Eksempler på slike forenlige formål er å levere en tilgjengelig og tilstrekkelig sikret tjeneste, samt levere opplæring, kundestøtte og forbedringer direkte knyttet til den aktuelle bruken. Grensen for hvilke formål som er forenelige med behandlingsansvarliges opprinnelige formål, og hvilke formål som faller utenfor er ofte vanskelig å avgjøre. Personvernforordningen oppstiller vurderingsmomenter i art. 6 nr. 4. Viderebehandling til trening av en KI-modell er et godt eksempel på en vanskelig vurdering, der både behandlingsansvarlig og databehandler har en interesse i at modellen er så god som mulig, i tråd med samfunns- og teknologiutvikling, og at tjenesten løpende ivaretar behandlingsansvarliges instruks etter artikkel 5 (blant annet konfidensialitet, integritet og tilgjengelighet). Slik behandling kan være i tråd med behandlingsansvarliges opprinnelige formål, og dermed ikke viderebehandling. Grensen kan dog være uklar og det kan være lett å miste oversikt og kontroll over behandlingen.

Samtidig kan det være vanskelig å se for seg at skoleeier kan tillate at personopplysninger om elever blir brukt til å trene tredjeparters maskinlæringsmodeller. Dette vil kunne medføre at databehandler selv blir behandlingsansvarlig, med egne formål, og dermed må ha et selvstendig behandlingsgrunnlag. Da vil opprinnelig behandlingsansvarlig miste retten til å utføre revisjon av behandlingsaktiviteter, som en ny behandlingsansvarlig igangsetter. Dette vil kunne ha implikasjoner for ivaretagelse av den registrertes rettigheter. Opprinnelig behandlingsansvarlig (skoleeier) kan ikke velge en databehandler som utfører slike behandlinger som selvstendig behandlingsansvarlig, uten å være sikker på at det foreligger et legitimt formål.

Det danske datatilsynet har i tre ulike avgjørelser fra januar 2024 skissert opp hvor grensene går for viderebehandling til utvikling av tjenester med blant annet kunstig intelligens (Datatilsynet, 2024a, Datatilsynet, 2024b). I de to første sakene er spørsmålet om et treningsdatasett kan deles videre. Vilkårene anses å være oppfylt i begge sakene, i den første fordi personopplysningene

allerede er allment og åpent tilgjengelige, og i den andre fordi det foreligger en kontrakt. I den tredje saken om Googles behandling av personopplysninger til egne (forretnings)formål, lander det danske Datatilsynet på at dette formålet ikke er forenlig og behandlingen ulovlig. Grensen trekkes ved behandling som innebærer utlevering til en annen behandlingsansvarlig for videreutvikling av deres tjenester.

Det er altså vesentlig både hva som er status på personopplysningene som brukes, hva som er behandlingsgrunnlag og eventuelt supplerende rettsgrunnlag, samt hvilket formål det utføres for. Kvalitetsutvikling i skolen er et formål med grunnlag i opplæringsloven, mens produktutvikling av kommersielle tjenester, ikke kan hjemles i samme bestemmelse – når det kommer til å bruke personopplysninger for å oppfylle formålet. En barriere i skolesektoren er at svært få personopplysninger kan ansees offentlig tilgjengelige eller at det foreligger et behandlingsgrunnlag som er egnet for viderebehandling, slik som oppfyllelse av avtale.

Forvaltningen skal ha kapasitet til å vurdere endringer og mekanismer for å hindre at nye behandlinger som ikke er i tråd med ansvarlighetsprinsippet igangsettes. Teknologiutviklingen generelt og kontinuerlig utvikling av den enkelte tjeneste vil medføre nye problemstillinger og behov for revisjon av vurdering. For skoleeier som behandlingsansvarlig er det viktig å kunne håndtere en slik økning i «forvaltningskostnaden».

6.3.1 Risiko og konsekvens ved viderebehandling

Tabell 7 - Risiko og konsekvens ved viderebehandling

ID	Risiko	Konsekvens	Mulig tiltak
6.3.01	At behandlingsansvarlig viderebehandler personopplysninger til nytt formål som potensielt ikke er forenlig med det opprinnelige formålet.	Eleven risikerer tap eller forringelse av rettigheter og friheter dersom opplysninger avgitt eller innsamlet til ett formål behandles for et annet.	Kompatibilitetstest jf. GDPR artikkel 6-4 for å påse at formålet med ny behandling av personopplysninger er forenlig med opprinnelig formål.
6.4.02	At databehandler viderebehandler personopplysninger for egne formål.	Eleven risikerer tap eller forringelse av rettigheter og friheter dersom opplysninger avgitt eller innsamlet til ett formål behandles for et annet.	-Implementere tiltak som sørger for at personopplysninger ikke inngår i data til KI-endepunktet. -Databehandler instrueres i å ikke viderebehandle inferensdata for egne formål. -Databehandleroppdraget avsluttes.

6.4 INTEGRASJON MED EKSTERNE DATAKILDER ELLER PROSESSER

Integrasjoner kan medføre flere ulike risikoer, men særlig tre er relativt unike til språkmodellenes funksjonalitet. Funksjonene til språkmodeller kan utvides ved å knytte de til eksterne datakilder eller prosesser, som kan kalles i tredjepartstjenester. Dette gjør det mulig å sende data til språkmodeller som ikke er i modellens minne, samt å bruke det modellen skaper til å instruere andre modeller eller tjenester. Det kan bidra til å øke presisjonen eller relevansen i responsen til språkmodellen. For eksempel, gjennom å instruere modellen til kun å utføre et søk i et dokument som inneholder et ordensreglement, heller enn å forutsette at det samme ordensreglementet er del av treningsdataene til modellen og kan gjenskapes i modellens respons. Ettersom flere ordensreglement eller ting som ligner på ordensreglement kan være identifisert gjennom trainingen som relevant, så er det tilnærmet umulig å vite hva modellen har lagt til grunn, med mindre disse aktiveringene kan avgrenses, samtidig som andre bevares. Dette kan gjøres ved å utvide det parametriske minnet til modellen, og en vanlig teknikk for dette er retrieval augmented generation, forkortet «RAG» (Lewis et al, 2020). RAG kan gjøre en kunnskapskilde tilgjengelig for en språkmodell, slik at den kan skape respons basert på denne kilden. RAG kan avhjelpe problemstillingen med at språkmodeller ikke presist kan representere fakta, gjennom å avgrense hva modellen har mulighet til å basere responsen på.

Overføring og behandling av data, begge veier, kan føre til betydelige risikoer. For data som hentes inn er det viktig å påse at kunnskapskilder er grundig gjennomgått og ikke inneholder opplysninger som modellen ikke skal kunne gjenskape. Prompt injection innebærer å overstyre modellens opprinnelige instruksjoner for å få den til å utføre operasjoner den ikke var ment å gjøre eller bevisst var trent for å ikke gjøre. Indirekte injisering av data (indirect prompt injection) søker å utføre det samme, men der bevisst manipulerende data gjøres tilgjengelig, typisk på internett, slik at modellen fanger det opp når den indekserer ressurser (Kai et al, 2023). Den kan i tillegg instrueres til å igangsette behandlingsaktiviteter gjennom tredjepartsressurser.

«We show how processing retrieved prompts can act as arbitrary code execution, manipulate the application's functionality, and control how and if other APIs are called.»
(Greshake et al, 2023)

For data som sendes ut er det flere problemstillinger. En av utfordringene med språkteknologi er at modellen ofte vil behandle et sett med inndata i flere omganger. Det vil si at opplysninger som ble gitt modellen i begynnelsen vil behandles gjentakende ganger for å gi modellen kontekst. Dersom utdata da ender opp med å behandles i en utilstrekkelig sikret fillagring, så kan det bidra til å sende personopplysninger til lokasjoner der de ikke skal behandles. Det er demonstrert relativt enkle teknikker for å trekke data ut fra både instruksjoner og eksterne datakilder (Burgess, 2023, Huang, 2022). En tjeneste som OpenAIs GPTs lar den som designer agenter, eller «bots», å bruke applikasjonsprogrammeringsgrensesnitt (APIer), som relativt enkelt og uten brukers kjennskap til det, ta opplastede personopplysninger og overføre til tredjeparter (Paterson, 2024).

Såkalt «function calling», å kalle funksjoner i eksterne prosesser/arbeidsflyter/verktøy gjør det mulig for en språkmodell å omsette språkhandlinger til funksjonelle instruksjoner i andre tjenester. Et eksempel på dette kan være å utføre en bestilling av flybilletter eller å styre en avatar i et

dataspill. Både «embedding»/vektorsøk og «function calling» kan ha implikasjoner ved behandling av personopplysninger, da det kan innebære en behandling på nye måter, i nye kontekster med nye behandlingsaktiviteter. Det er i tillegg demonstrert relativt enkle teknikker for å trekke data ut fra både instruksjer og eksterne datakilder (Burgess, 2023, Huang, 2022). En tjeneste som OpenAIs GPTs lar den som designer agenter, eller «bots», å bruke applikasjonsprogrammeringsgrensesnitt (APIer), som relativt enkelt og uten brukerens kjennskap til det, ta opplastede personopplysninger og overføre til tredjeparter (Paterson, 2024).

Denne type risiko kan ha betydelige konsekvenser for personopplysningsvernet da formålet med å bryte ned sikkerhetsmekanismer som er bygget rundt språkmodellene kan medføre utilsiktede eller tilsiktede, men ulovlige, behandlingsaktiviteter. Særlig risiko der tjenesten er tilknyttet eksterne prosesser som kan medføre at en behandlingsaktivitet forplanter seg gjennom flere tjenester (Harang, 2023). Det er slikt sett viktig og relevant å vurdere hvilke datakilder som legges til grunn, samt hvilke tredjeparter som har tilgang til å behandle data fra en tjeneste, og ivaretagelse av personopplysningssikkerhet jf. personvernforordningen artikkel 5 nr. 1 bokstav f og art. 32.

6.4.1 Risiko og konsekvens ved integrasjon med eksterne datakilder eller prosesser

Tabell 8 - Risiko og konsekvens ved integrasjon med eksterne datakilder eller prosesser

ID	Risiko	Konsekvens	Mulig tiltak
6.4.01	<p>Manglende oversikt og kontroll over dataflyt mellom komponenter, inkl. håndtering av inndata behandlet i KI-løsning i strid med GDPR artikkel 5, annet aktuelt lovverk eller internrutine.</p> <p>Personopplysninger på avveie, for eksempel ved dataeksfiltrasjon ved bruk av blant annet RAG-løsninger, slik som OpenAI GPTs utviklet av tredjeparter. Utilsiktet behandling kan også forekomme i slik som sideoppsummering i Microsoft Copilot i Edge.</p> <p>Utilstrekkelig sikkerhet ved autentisering, overføring eller lagring.</p>	<p>Manglende kontroll på behandlingsaktiviteter og personopplysninger kan medføre høy risiko for brudd på flere personvernprinsipper.</p> <p>Lovlighet: behandling uten hjemmel.</p> <p>Dataminimering: flere opplysninger enn nødvendig.</p> <p>Konfidensialitet: behandling for konfidensielle eller taushetsbelagte opplysninger, samt opplysninger underlagt sikkerhetsbestemmelser.</p>	<p>Implementere rutiner for å ha kontroll på komponenter i et KI-system, inkl. overføringer, lagring og tilganger, foretatt både av systemer og personer.</p>
6.4.02	<p>Utilstrekkelig sikkerhetsnivå ved autentisering til tjeneste, administrasjonsløsning eller back-end, i strid med GDPR artikkel 32.</p>	<p>Tredjeparter kan potensielt oppnå tilgang til løsningen eller komponentene, og injisere kode i eller omkonfigurere disse.</p>	<p>Jevnlig vurdere om sikkerhetsnivå er tilstrekkelig for å ivareta personvernprinsippene i disse.</p>

			stort, inkl. men ikke uttømmende: <ul style="list-style-type: none"> - Konfidensialitet og integritet - Formålsbegrensning - Riktighet
6.4.04	Feil ved tilgangsstyring som gir for høye rettigheter i eksterne komponenter. Ikke i tråd med prinsippet om «Least privileged access», og i strid med GDPR artikkel 32, forskrift til Opplæringslova §22A-2.	Brudd på konfidensialitet Person uten tjenstlig behov kan få tilgang til opplysningene og kan potensielt viderebehandle de til nye formål.	Implementere rutiner for å ivareta prinsipper og rutiner om tilgang til å behandle personopplysninger, fra kilde system til læringsressurser og fagsystem.

6.5 REIDENTIFISERING AV PSEUDONYMISERTE ELLER ANONYMISERTE OPPLYSNINGER

En språkmodell som er trent på personopplysninger, enten det er i åpent tilgjengelige webdata eller andre åpne kilder, eller som er trent på eller rettet mot data i skoleeiers egne systemer eller hos leverandører, kan i ulike tilfeller utlede slutninger som reidentifiserer personer. Dette kan for eksempel forekomme dersom instruksene til språkmodellen inneholder opplysninger som leder modellen til å utlede at andre opplysninger mest sannsynlig skal inkluderes i det den returnerer i utdata (Staab et al, 2023). Konseptet demonstreres gjennom <https://llm-privacy.org>. Dette kan være en særlig risiko når det kommer til å utlede lokalisering opplysninger om personer på adressesperretiltak.

Det mest åpenbare tiltaket er å begrense hvilke opplysninger en språkmodell har tilgang til, både det som utgjør treningsdata, som er eksemplene modellen trenes på, samt det som gis av ulike inndata, som kan inneholde personopplysninger. Dette er data som kan påvirke modellens vektning, og i gitte tilfeller gjenskapes i responsen. Mindre modeller trener færre parametere som potensielt kan gjenskape personopplysninger. Det bør ikke tas i bruk språkmodeller basert på treningsdatasett som ikke er tilstrekkelig «vasket» for personopplysninger. Dette kan være vanskelig å teste for. Særlig der treningsdatasett ikke er kjent. Det kan dessuten være vanskelig å forutsi hvilke opplysninger som kan utgjøre personopplysninger eller være lokalisering i en viss kontekst. Det viktigste tiltaket kan derfor være å hindre språkmodeller i å behandle opplysninger i applikasjoner og nettsider, slik som skoleadministrative system, pedagogiske fagsystem og digitale læringsressurser. Tester utført med Microsoft Copilot i nettleseren Edge indikerer at det behandles personopplysninger i nettinnholdet når tjenesten får tilgang til det, og at disse opplysningene behandles i modellen og potensielt viderebehandles i indekseringen i søkemotoren. Microsoft opplyser at det ikke påvirker rangering av søkeresultater, men det vil fremdeles forekomme en behandling (SLM Rijk, 2024).

I eksempelet under har Copilot i Edge lest inn navnet på personen fra en nyhetsartikkel på et intranett. Instruksene til modellen er at brukeren er nødt til å gi tilbakemelding på artikkelen som er skrevet, og at den må sendes på epost til vedkommende. Navn og epost til artikkelforfatteren

er anonymisert. «ContextData» er inndata som sendes modellen, som også består av tidligere skapt utdata:

«contextData.txt

```
{ "prompt": "Hvordan kan jeg følge opp med NAVN?", "promptId": "xxxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxx", "responseMessage": "For å følge opp med NAVN, kan du gjøre følgende:\n\n1. **Send en e-post:** Bruk e-postadressen du har (NAVN.NAVNESEN@SKOLEEIER.no) for å sende en oppfølgingsmelding. Du kan spørre om han har mottatt tilbakemeldingen og om han har noen spørsmål eller kommentarer.\n\n2. **Personlig kontakt:** Hvis du ser NAVN på skolen, kan du ta en prat med ham og spørre om han har fått tilbakemeldingen og om han har noen spørsmål.\n\n3. **Bruk skolens kommunikasjonsverktøy:** Mange skoler har interne kommunikasjonsplattformer som Teams, Google Classroom eller lignende. Du kan bruke disse verktøyene til å sende en melding og følge opp.\n\nHer er et eksempel på en oppfølgingsmelding du kan sende:\n\n---\n\nHei NAVN,\n\nJeg ville bare følge opp og høre om du har mottatt tilbakemeldingen jeg sendte på artikkelen din. Har du noen spørsmål eller kommentarer til tilbakemeldingene?\n\nVennlig hilsen, [Navnet ditt]\n\n---\n\nLykke til! Hvis du trenger mer hjelp, er det bare å spørre. 😊", "chatHistory": [ { "message": "Hvordan kan jeg følge opp med NAVN?", "user": "user" }, { "message": "For å følge opp med NAVN, kan du gjøre følgende:\n\n1. **Send en e-post:** Bruk e-postadressen du har (NAVN.NAVNESEN@SKOLEEIER.no) for å sende en oppfølgingsmelding. Du kan spørre om han har mottatt tilbakemeldingen og om han har noen spørsmål eller kommentarer.\n\n2. **Personlig kontakt:** Hvis du ser NAVN på skolen, kan du ta en prat med ham og spørre om han har fått tilbakemeldingen og om han har noen spørsmål.\n\n3. **Bruk skolens kommunikasjonsverktøy:** Mange skoler har interne kommunikasjonsplattformer som Teams, Google Classroom eller lignende. Du kan bruke disse verktøyene til å sende en melding og følge opp.\n\nHer er et eksempel på en oppfølgingsmelding du kan sende:\n\n---\n\nHei NAVN,\n\nJeg ville bare følge opp og høre om du har mottatt tilbakemeldingen jeg sendte på artikkelen din. Har du noen spørsmål eller kommentarer til tilbakemeldingene?\n\nVennlig hilsen, [Navnet ditt]\n\n---\n\nLykke til! Hvis du trenger mer hjelp, er det bare å spørre. 😊", "user": "bot" } ] }
```



6.5.1 Risiko og konsekvens ved reidentifisering

Tabell 9 - Risiko og konsekvens ved reidentifisering

ID	Risiko	Konsekvens	Mulig tiltak
6.5.01	Reidentifisering av person, der kunstig intelligens blir brukt f.eks. til å saksbehandle dokumenter der personens navn eller andre identifiserende opplysninger er trukket ut eller pseudonymisert.	Potensielt brudd på konfidensialitet og taushetsplikt, samt formålsbegrensning.	Tilstrekkelig kontroll på treningsdata og eksterne kunnskapskilder som kan inneholde personopplysninger. Mekanismer for å hindre modellen i å generere utdata som kan reidentifisere. Avgrensning av mulighetsrom for aktivering av parametre i modellen. Enten gjennom kontekstualisering av parametrisk minne eller embedding. Ikke bruke tjenester som kan analysere nettsideinnhold i nettleser og skape sammenstillinger ved bruk av en språkmodell.

6.6 MISVISENDE, VILLEDENDE ELLER SKADELIG INNHOLD

KI-systemer bygget på generativ kunstig intelligens, slik som Chat-GPT, har blitt enormt populære det siste året, også for elever i skolen. Det har ført med seg fordeler, men også flere risikoer, blant annet for at systemet genererer feilaktig eller misvisende informasjon. Dette kan ha konsekvenser for elevenes forståelse av verden og evne til å ta informerte beslutninger. Videre er det en risiko for eksponering for upassende innhold eller samtaler som kan ha en negativ innvirkning på et barn og unges utvikling. Det foreligger ingen enhetlig definisjon av innholdstyper, men kritisk medieforståelse defineres blant annet gjennom kompetanse til å vurdere i hvilken grad noe er misvisende (Medietilsynet, 2021). Heller ikke Digital Services Act (DSA, 2022/2065), som regulerer nettbasert innhold, definerer innholdstypene – kun håndhevingen. Det kan ansees som en nødvendig skjønnsmessig vurdering, i hvilken grad noe er misvisende, villedende eller skadelig. Slikt innhold er i seg selv heller ikke ulovlig innhold, og må sees i sammenheng med annet aktuelt lovverk, jf. artikkel 2 i DSA. Filtrering av innhold er et aktuelt og i enkelte tilfeller effektivt tiltak, men det er samtidig en utfordring å definere hvilke kategorier et filter skal fange, uten å begrense en fri tilgang til informasjon som ikke i seg selv er ulovlig. Det er komplisert og potensielt problematisk å definere for mange kategorier og for bredt. For tilganger til netttinnhold via domenespesifikke adresser, så kan det legges til grunn et nettverksfilter, slik som [Internettfilter – Politiet.no](https://www.politi.no/interneffilter) (Kripos-filter). Et slikt filter vil vel og merke ikke nødvendigvis begrense innhold i teknologi som integreres gjennom et API, med mindre domenet det kaller også blokkeres. Mange API kalles gjennom vanlige skyplattformdomener som brukes i andre tjenester, noe som gjør blokkering vanskelig.

Generering av innhold i språkmodeller fungerer ikke på samme måte som et tradisjonelt informasjonssøk, og er ikke nødvendigvis avhengig av å navigere til nettadresser som kan sperres. Responsen til språkmodeller må modereres på språknivå, gjennom å regulere hvilke kommunikative handlinger den kan foreta (Cohen & Levesque, 1995 og Castelfranchi, 1998). Det er samtidig mulig å tenke seg en regulering av hvilke kilder en språkmodell kan kalle for berikelse av resultater gjennom metoder slik som «RAG» eller «function calling». Det er f.eks. viktig å hindre veiledning i tilvirkning av og eksponering til potensielt skadelige og farlig kjemiske stoffer, samt våpen. Moderasjon kan forsøke å forhindre dette, men kan i enkelte tilfeller omgås ved å endre språklig argumentasjon i instruksene til språkmodellen. Det finnes noen ulike teknikker for å avgrense hva en språkmodell vil generere av respons og hva responsen er basert på. I fagfeltet og i faglitteraturen omtales dette samlet sett som «guardrails», der formålet er å sørge for at kunstig intelligens ikke handler på tvers av definerte mål og verdier – det som på KI-feltet betegnes som «alignment».

Språkmodeller vil generere en respons basert på vekt og bias, som indikerer styrkene i forbindelse mellom lærte funksjoner i modellen. Ulike teknikker for videretrening kan benyttes for å begrense mulighetene en språkmodell har for f.eks. å skape innhold som bryter med det ønskede verdigrunnlaget i konteksten den tas i bruk. Dette kan være effektive midler, men har konsekvensen at den legger stor grad av definisjonsmakt i hendene på den aktøren som definerer de. Dette kan være uønsket, særlig der det er manglende åpenhet rundt slike avgrensninger eller regelsett. Utfordring med moderasjon er at den som utformer reglene sitter med definisjonsmakten. Wikipedia forsøker å adressere dette gjennom en konsensusstilnærming og vedlikeholder en liste med kilder og gradering av pålitelighet (Wikipedia, 2024), men en prosess som også utfordres av uklarheter rundt hva og i hvilken grad innhold er skapt med KI (Edwards, 2024a). Det kan også føre til uheldige situasjoner, i forhold til at optimalisering for en effekt har en utilsiktet eller uønsket effekt for en annen (Edwards, 2024b).

Generativ kunstig intelligens er en relativt ny teknologi og det er uklart per d.d. om bruk av språkmodeller til å skape innhold som typisk regnes som ulovlig å tilvirke i fysisk forstand eller som kan regnes som ulovlig i for eksempel USA, vil kunne falle inn under kategorier av data som forbindes med lover med ekstraterritoriell jurisdiksjon, slik som Section 702 av FISA og CLOUD-Act. Det er også uklart i hvilken grad amerikanske leverandører, slik som Microsoft, OpenAI, Google og Amazon, men også norsk selskaper som tilbyr tjenestene sine i USA, vil måtte etterleve en begjæring eller forespørsel om utlevering av opplysninger eller merbevaring (GreenbergTraurig, 2022). De fleste applikasjonsprogrammeringsgrensesnitt (API), fordrer moderasjon med leverandørens standardinnstillinger, der det enten ikke er mulig å gå lavere enn standardverdiene eller at moderasjon må skrues av, med den konsekvens at leverandøren frasier seg ansvar for tjenestens kapabiliteter til å generere denne type innhold.

6.6.1 Risiko og konsekvens ved misvisende, villedende eller skadelig innhold

Tabell 10 - Risiko og konsekvens ved misvisende, villedende eller skadelig innhold

ID	Risiko	Konsekvens	Mulig tiltak
6.6.01	Mangelfull eller feilkonfigurert moderasjon av inn- og utdata som behandles av KI-tjeneste.	KI-tjenesten vil kunne utføre behandlinger som strider med intensjonen til behandlingsansvarlig. F.eks. produksjon av skadelig multimodalt innhold eller utføre skadelig funksjonskall til tredjepartstjenester.	Jevnlig teste og revidere («Red-teaming») av KI-system for å avdekke feilkilder. Avgrense mulighetene KI-systemet har til å viderekommunisere utdata. F.eks. begrense tilgjengelige API.

7 BEHANDLINGSAKTIVITETER

Behandlingsansvarlig må definere hovedrettsgrunnlag etter artikkel 6 i personvernforordningen, og i tråd med opplæringslova §25-1.

Tabell 5 - Behandlingsaktiviteter

ID	Behandlingsaktivitet	Formål og forslag til supplerende rettsgrunnlag
01.01	Bruk av kunstig intelligens til å analysere multimodalt innhold i elevens produksjon og skape oppsummeringer/sammendrag og anbefalinger til eleven selv.	Bidra til å oppfylle retten til underveisvurdering, gjennom å tilpasse opplæringen for økt læringsutbytte. Opplysninger om elevens læring, læringsanalyse Opplæringsloven §25-1 (Behandling av personopplysninger) jf. opplæringsforskrifta §9-11 (Underveisvurdering).
01.02	Bruk av kunstig intelligens til å analysere multimodalt innhold i elevens produksjon og skape sammendrag og anbefalinger til elevens lærer.	Analysere multimodalt innhold for å gi lærer indikatorer på hva eleven har vist av kompetanse i sin produksjon. Opplæringsloven §25-1 (Behandling av personopplysninger) jf. opplæringsforskrifta §9-11 (Underveisvurdering).
01.03	Bruk av kunstig intelligens til å analysere multimodalt innhold i elevgruppens produksjon og skape sammendrag og anbefalinger til elevgruppens lærer.	Analysere multimodalt innhold for å gi lærer indikatorer på hva elevgruppen har vist av kompetanse gjennom sin produksjon. Opplæringsloven §25-1 (Behandling av personopplysninger) jf. opplæringsforskrifta §9-11 (Underveisvurdering).
02.01	Automatisert innholdsfiltrering (content filtering)	Formål: Begrense at elever får tilgang til skadelig innhold på nett og i bruk av skoleverktøy. Primært behandlingsgrunnlag: GDPR artikkel 6(1) c) Rettslig forpliktelse Supplerende behandlingsgrunnlag: Opplæringsloven § 25-1 (Behandling av personopplysninger), jf. § 12-3 (Nulltoleranse og førebyggjande arbeid) Bildeprogramloven § 8 (Plikt til å beskytte mindreårige mot skadelig innhold) GDPR artikkel 32(1) (Plikt til å ivareta sikkerhet ved behandling av personopplysninger) og

		eForvaltningsforskriften § 15(2) (Plikt for forvaltningsorganer til å ha internkontroll for informasjonssikkerhet)
02.02	Automatisert overvåkning og logging av tjenestemisbruk (abuse monitoring)	<p>Formål: Ivareta sikkerhet og redusere sannsynligheten for misbruk av tjenester.</p> <p>Primært behandlingsgrunnlag: GDPR artikkel 6(1) e) Oppgave i allmennhetens interesse eller utøvelse av offentlig myndighet.</p> <p>Supplerende behandlingsgrunnlag: GDPR artikkel 32(1) (Plikt til å ivareta sikkerhet ved behandling av personopplysninger)</p> <p>og</p> <p>eForvaltningsforskriften § 15(2) (Plikt for forvaltningsorganer til å ha internkontroll for informasjonssikkerhet)</p>
02.03	Menneskelig behandling av loggført tjenestemisbruk (human review of abuse monitoring)	<p>Hvilket grunnlag for brudd på privatrettslige tjenestevilkår?</p> <p>Formål: Gjennomføring av internkontroll, følge opp potensielt misbruk av tjenester, sikkerhetshendelser og andre uønskede hendelser.</p> <p>Primært behandlingsgrunnlag: GDPR artikkel 6(1) e) Oppgave i allmennhetens interesse eller utøvelse av offentlig myndighet.</p> <p>Supplerende behandlingsgrunnlag: GDPR artikkel 32(1) (Plikt til å ivareta sikkerhet ved behandling av personopplysninger)</p> <p>og</p> <p>eForvaltningsforskriften § 15(2) (Plikt for forvaltningsorganer til å ha internkontroll for informasjonssikkerhet)</p> <p>Formål: Sikre et godt skolemiljø gjennom å oppfylle skoleeiers aktivitetsplikt i en potensiell skolemiljøsak.</p> <p>Primært behandlingsgrunnlag: GDPR artikkel 6(1) c) Rettslig forpliktelse og artikkel 9(2) g) Viktige allmenne interesser</p> <p>Supplerende behandlingsgrunnlag: Opplæringsloven § 25-1 (Behandling av personopplysninger), jf. Opplæringsloven kapittel 12 (Skolemiljøet til elevene - reglene om skolens aktivitetsplikt og undersøkelsesplikt)</p>

8 ANBEFALTE TILTAK FOR SKOLEEIER

Skoleeiers tiltak må vurderes i lys av omfanget av behandlinger som det planlegges for lokalt, i digitale læringsressurser og pedagogiske fagsystem. Selv om det er gjort overordnede vurderinger og foreslåtte tiltak, samt foreslåtte formål og behandlingsaktiviteter, så kan ikke denne vurderingen tas til inntekt som en instruks til den enkelte skoleeier, som er selvstendig behandlingsansvarlig. Det er derfor viktig at hver skoleeier selv gjennomgår alle vurderinger, foreslåtte tiltak, formål og behandlingsaktiviteter, og selv kun igangsetter de behandlinger som den selv vurderer å kunne ansvarlig forvalte.

Det drøftes og foreslås her tiltak for å redusere konsekvenser for den registrerte, for hver av personvernprinsippene, jf. artikkel 5 i personvernforordningen (GDPR).

8.1 LOVLIGHET

Hvorvidt behandlingen av personopplysninger er lovlig er betinget av at de er vurdert og dokumentert i henhold til prinsippene i artikkel 5. Skoleeier skal sørge for at læringsressurser med KI ikke tas i bruk før det er gjort en vurdering av tjenesten. Ettersom det er stadig økt tilfang av digitale tjenester med KI, der mange er enkelt tilgjengelig via nett, så er en vesentlig del av skoleeiers lovlighetsvurdering betinget av rutiner innrettet overfor virksomhetene. Disse skal være egnet for å fange opp behov og anskaffelser i virksomhetene, samt prosesser og ressurser for å håndtere dette. De som anskaffer KI-tjenester må kjenne til hva som kreves for at de *lovlig* skal kunne ta i bruk, og igangsette en behandling, i en digital tjeneste.

I tillegg til etterlevelse av de øvrige prinsippene, er lovligheten av en behandlingsaktivitet også betinget av at det foreligger et behandlingsgrunnlag, samt et supplerende rettsgrunnlag der det er nødvendig. Personvernforordningen er teknologinøytral og gjelder både automatisert og manuell behandling av personopplysninger, enten i strukturert eller ustrukturert form (European Commission, 15.04.2024). Til tross for dette, så vil behandlingsaktiviteter utført ved bruk av enkelte teknologier og datasett, være uforholdsmessig, ikke nødvendig, og potensielt utgjøre en høy risiko for den registrertes rettigheter og friheter.

Personvernforordningen art. 6 nr. 1 oppstiller seks lovlige behandlingsgrunnlag. Det er art. 6 nr. 1 bokstav c og e som er mest relevante for bruk av KI i læringsressurser. Ekspertgruppen i NOU 2023:19, «Læring, hvor ble det av det i alt mylderet», drøfter de to rettslige grunnlagene og legger til grunn at bokstav e, behandlingen er nødvendig for å utføre en oppgave i allmennhetens interesse eller utøve offentlig myndighet som den behandlingsansvarlige er pålagt, er mest relevant.¹⁰ Etter art. 6 nr. 3 kreves det et supplerende rettslig grunnlag i nasjonal rett eller unionsretten som den behandlingsansvarlige er underlagt. Relevante supplerende rettsgrunnlag kan finnes i opplæringsloven, forskrift til opplæringsloven og læreplanverket. Dersom det skal behandles særlige kategorier av personopplysninger er det viktig at

¹⁰ Se punkt 10.2.1.

behandlingsansvarlige påser at den har lovlig grunnlag for behandling av nevnte kategorier personopplysninger jf. artikkel 9 i personvernforordningen (GDPR).

Ekspertgruppen til NOU 2023:19 gjør en nærmere vurdering av opplæringsloven §§ 1-3 (ny lov § 11-1) og 13-3e, og forskrift til opplæringsloven § 3-2, men finner at ingen av disse bestemmelsene er egnet til å presisere det rettslige grunnlaget for behandling av personopplysninger i læringsanalyse. Den generelle bestemmelsen i opplæringsloven § 15-10 som understreker at skoleeier kan behandle personopplysninger når det er nødvendig for å utføre en oppgave i loven er også omtalt¹¹:

Å konstatere at læringsanalyse kan være i samsvar med de overordnede formålene i loven, er ikke nødvendigvis så utfordrende. Men det er ikke nok å konstatere samsvar. Som vist i kapittel 5.2.1 stiller personvernforordningen og EMK-retten krav til hvordan det rettslige grunnlaget for å behandle personopplysninger er utformet. Det rettslige grunnlaget «must lay down clear and precise rules governing the scope and application of the measure in question» (C-439/19 Latvias Republikas Saeima [Storkammer], 2021, avsnitt 105).

NOUen drøfter også hvorvidt et supplerende rettsgrunnlag for en maskinell behandling av personopplysninger finnes i opplæringslova, slik den fremstår i dag. Ekspertgruppen peker særlig på krav til nødvendighet og forholdsmessighet ved en behandling, samt at behandlinger som kan utgjøre et inngrep i EMK artikkel eller Grunnlovens §104 må ha et tydelig og presist grunnlag, være forutsigbart for den registrerte – samt for øvrig i tråd med artikkel 5 (NOU2023:19, 2023). Kunnskapsdepartementet har per d.d. ikke offentliggjort sin innstilling til NOU 2023:19, slik at synspunktet til lovgiver er uavklart og om det vurderes en konkret hjemmel for maskinell behandling av personopplysninger.

Mangelen på en nærmere avgrensning av behandlingsaktiviteter der KI utgjør del av midlene tilsier at skoleeier som behandlingsansvarlig, selv må gjøre nødvendige avgrensninger, slik at disse er i tråd med GDPR. Lovlighetsspørsmålet handler ikke utelukkende om å identifisere et behandlingsgrunnlag, men en etterlevelse av personvernprinsippene i sin helhet. Ettersom det her er snakk om ny og dels ukjent teknologi og praksis og et stort antall sårbare registrerte, så stiller det særlig krav til å utlede lovligheten fra vurderinger om etterlevelse av de øvrige prinsippene. I et brev datert 19. juni 2024 [off.], presiserer Kunnskapsdepartementet at opplæringslova §25-1 ikke gir grunnlag for viderebehandling av personopplysninger for andre formål enn de som direkte følger av oppgaver som utføres etter loven, opplæringsforskrifta eller læreplanverket. Videre presiserer departementet at overordnet del av læreplanverket anses som for overordnet til å utgjøre et supplerende rettsgrunnlag for overordnede behandlingsaktiviteter som ikke er direkte knyttet til lovpålagte oppgaver. Det er dermed viktig at behandlingsaktiviteter grunnes så konkret som mulig i oppgaver knyttet til lov og forskrift – for opplæringsens tilfelle vil dette innebære at enkelte behandlingsaktiviteter grunnes i den enkelte læreplan.

- Utforme tydelige formål, grunnet i oppgaver som skal utføres etter lov og forskrift.

¹¹ Se punkt 10.3.5.

- Drøft og vurder nødvendigheten av behandlingene som utføres for å oppfylle formålet.

8.2 RETTFERDIGHET

Om en behandling av personopplysninger med kunstig intelligens skal være rettferdig er det nødvendig å vurdere om KI-tjenesten og modellene den bygger på, innehar bias fra treningsdata, trening, eller implementering og brukskontekst, og at utdata som er påvirket av disse faktorene, tillegges betydning og legges til grunn av den registrerte selv, eller av personer eller andre system som behandler disse dataene. Rettferdighetsprinsippet er tett forbundet til den grunnleggende rett til ikke-diskriminering.

En av driverne bak KI-tjenester rettet mot sluttbrukere er å effektivisere økningen i graden av brukertilpasning. Nettjenester har gått fra å være én tjeneste til alle, og til i økende grad å bli brukertilpasset på tidlig 2000-tallet gjennom designtilpasninger, profilering, og sporing. Kunstig intelligens tilbyr metoder for å ytterligere tilpasse systemenes respons til brukeren ved å trekke på enda flere datakilder, på en langt mer effektiv måte. Der tidligere én autoritativ kilde ga økt tillit, kan KI-tjenester med lavere grad av tilpasning risikere å fremstå som kvalitativt sett dårligere tjenester.

- Være bevisst hvilke data som ligger til grunn for utvikling av KI-tjenesten, samt hvilke eksterne data den har tilgang på ved bruk.
- Teste og validere respons fra KI-tjenesten, for å avdekke om det er indikasjoner eller tydelig grunnlag for at den skaper respons som innehar skjevheter eller feil.

8.3 ÅPENHET

KI-systemer består av et relativt stort antall teknologiske komponenter som er vanskelig å få oversikt over og innsyn i, noe som gjør det vanskelig å forklare på en enkel måte. Det er ikke mulig for brukere av tjenester med kunstig intelligens (slik som ChatGPT), å forklare modellens oppførsel (Huang et al, 2022). Hvordan skiller tjenesten resultatene fra en syntese av tekst, fra ulike treningsdata?

Uønsket eller uhensiktsmessig vekting og bias, fra interaksjon med språkmodeller, kan forsterkes gjennom lengre kontekster/samtaler, samt caching/minne på tvers av samtaler. Et fenomen som observeres både i maskinlæring, lingvistikk og samfunnsvitenskapen, og med relevans for interaksjon med og anvendelse av kunstig intelligens er konseptet «stiavhengighet, og faktorer som medfører økte kostnader eller fravær av mulighet til å endre sti (Grube, 2015).

Utilstrekkelig åpenhet rundt behandlingene, og hva som ligger bak resultatene i en instruks, kan føre til at det blir vanskelig for den registrerte å forstå, å utøve sine rettigheter, samt potensielt også mer usikkerhet og konflikt rundt nødvendigheten. Dette kan medføre risikoer knyttet til åpenhet.

- Være bevisst om hvilken underliggende KI-teknologi (maskinlæringsmodeller), som implementeres i digitale læringsressurser og fagsystem, og helst implementere modeller med åpne vekter, samt åpne treningsdata.
- Gi god informasjon om KI-teknologi og skytjenester som brukes i behandling av personopplysninger.

8.4 FORMÅLSBEGRENSNING

Personopplysninger skal kun behandles for «spesifikke, uttrykkelig angitte og berettigede formål og ikke viderebehandles på en måte som er uforenlig med disse formålene» jf. GDPR artikkel 5-1 bokstav b. Det er ulike årsaker til at behandlinger med kunstig intelligens har en særlig risiko for å medføre brudd på prinsippet om formålsbegrensning.

1. Kunstig intelligens kommer i mange tilfeller i en form der teknologien kan anvendes for mange ulike formål. Det kan i enkelte tilfeller være relativt lav terskel for at personopplysninger behandles til nye formål, som ikke er forenlig med det opprinnelige.
2. Inn- og utdata behandlet i KI-tjenester lagres i mange tilfeller i kortere eller lengre perioder enn det som er nødvendig for å utføre inferens (å utlede nye setninger fra forutgående). Logging av inn- og utdata er vanlig praksis i KI-tjenester; data som er ettertraktet i modelltrening eller finjustering.

Det at opplysningene er tilgjengelig i en for videretrening av den underliggende modellen eller videreutvikling av andre tilknyttede tjenester.

Risikoer for at personopplysninger som behandles i tjenesten kan bli behandlet for andre formål kan reduseres gjennom juridiske, tekniske og organisatoriske tiltak:

- Opplæring i mulige risikoer og tiltak for brudd på prinsippet om formålsbegrensning.
- Beskrivelser av risikoer innenfor vurdering, faglig utvikling, og skoleadministrasjon.
- Databehandleravtale som tydelig regulerer behandlingen av personopplysninger i KI-systemer. Påse at avtalen tydelig definerer formål og omfang (midler).
- Avklare hvorvidt inndata behandles for å gi mer relevante svar fra modellen, og hvorvidt disse opplysningene lagres i en periode for utviklingsformål.

8.5 DATAMINIMERING

Dataminimering er prinsippet om å avgrense behandlingen av personopplysninger til det som er nødvendig for å oppfylle formålet. I enkelte læringssituasjoner med enkelte læringsressurser, kan det være en utfordring å avgrense hvilke personopplysninger som er nødvendig eller kan bli del av en behandling. Læring i en del sammenhenger, faglig eller sosialt, kan i mange tilfeller kreve at eleven har en utforskende tilnærming og drøfter egen livsverden. Det kan da også være nødvendig for eleven å forfatte det i en tekst, og for lærer å gi eleven veiledning. Ved bruk av læringsressurser med kunstig intelligens kan det derfor være en utfordring å avgrense helt hvilke opplysninger som behandles. Læreren kan avgrense oppgaven, men det er et vesentlig

prinsipp at eleven selv skal bidra til å definere hvordan deres læring skal foregå og hva som inngår i den. Heller enn å skulle begrense bruk av fritekst og brukergenerert innhold, som kan føre til svært uønskede effekter for undervisning og læring, så er det viktig å være bevisst på hvilke tjenester som kan brukes til hva, og at læring som kan føre til registrering av opplysninger som må vernes særskilt, ikke utføres i læringsressurser der behandlingene ikke i tilstrekkelig grad kan minimeres.

- Gi god informasjon om viktigheten av å være bevisst på hvilke opplysninger som registreres i en digital læringsressurs eller fagsystem med kunstig intelligens.
- Utarbeid rutiner for å ta opp prinsippet om dataminimering og tiltak for å ivareta dette, både internt og i leverandørpølfølgning.

8.6 RIKTIGHET

Implikasjonene for riktighet vil variere avhengig av behandlingsaktivitetene, og teknologien som brukes for å utføre disse. Ettersom det er såpass mange ulike måter responsen utformes til brukeren, og den i de fleste tilfeller ikke vil være i stand til å vurdere hva som er skapt på hvilket grunnlag, så må det legges til grunn at det er riktigheten i utdataene, uavhengig av kilde, som er vesentlig.

En vanlig innstilling for økt eller mindre variasjon i responsen til en språkmodell er det som betegnes som temperatur. Det trekkes ofte paralleller til begrepet kreativitet, som en menneskelig kognitiv prosess, for å gjøre det forklarbart. En variabel typisk mellom 0.0 og 2.0, en parameterkontroll som instruerer modellen i å prioritere hvilke tokens som skal benyttes i en respons. Dette varierer graden av usikkerhet/tilfeldighet som introduseres i sannsynlighetsberegningen, og har en direkte konsekvens på reliabilitet fordi det gradvis blir vanskeligere å reprodusere resultatet ettersom temperatur økes. Høy temperatur er også forbundet med høyere grad av konfabulasjoner i modellens respons. Temperaturinnstillinger kan være aktuelt tiltak i oppsett av KI-tjenester for å motvirke uønskede og utilsiktede effekter av variasjon, men må ikke kun sees opp mot type system, men hva som er formålet. For eksempel så vil et fagsystem for å understøtte underveisvurdering ikke nødvendigvis ønske høy variasjon, mens et fagsystem for å understøtte planlegging av undervisning kan ønske dette for å gi flere typer eksempler. Likeså, en læringsressurs som skal understøtte elevens læring om historiske fakta og presise beregninger ønsker muligens lavere variasjon, mens idégenerering og språklek ønsker høyere grad.

Opplysninger som sendes inn kan være riktig data, men variasjonen som språkmodellen introduserer, samt tilleggsmekanismer slik som moderasjon/filtrering eller eksterne datakilder, kan føre til at den genererte responsen ikke svarer til kriteriet for riktighet. Riktighet vil være av særlig konsekvens for den registrerte i formål og behandlingsaktiviteter som er av vesentlig betydning for den registrerte, og der det for etterrettelighetens skyld er viktig å kunne demonstrere at opplysningene som behandles er de vi tror (validitet) og at responsen må kunne reproduseres (reliabilitet). Dette er av svært vesentlig betydning i beslutningsstøttesystemer, noe vesentlig i innhold som skal fremstå som tillitsfull, og mindre vesentlig når det skapes noe som forkastes kort tid etterpå. Utdfordringen er å klassifisere en slik behandling i forkant, og å kunne forutse dette. Her kan f.eks. pedagogiske fagsystemer, som lærer benytter som støtte til

undervisvurdering, og læringsressurser som elever benytter for å fremme egen læring, oppstilles på hver sin side av en slik akse.

Bruk av generativ kunstig intelligens til å skape faglige vurderinger eller innhold som forventes å være presist, kan by på utfordringer med dagens teknologi. Dette er blant annet en særlig problemstilling der det er knyttet til styringsinstrumenter i skolen slik som læreplanene. Disse er detaljbeskrevet¹² og er gjenstand for en lengre prosess med mange aktører, der små språklige nyanser tillegges stor viktighet. Det betyr ikke at det ikke skal tolereres variasjon i begrepsbruken, slik personer omtaler og bruker disse. Det vesentlige er at en maskinell behandling må forventes å ha et presist utgangspunkt, blant annet for å ivareta etterrettelighet og kontinuitet i en rekke av behandlinger som kan resultere i at sluttresultatet har et betydelig endret innhold. Det er per i dag uklart hvilken grad en språkmodell som skaper nytt innhold basert på et stort antall varierte eksempler gitt i et stort treningsdatasett, vil kunne oppnå en høy grad av nøyaktighet. Tester av begrep innen tema læringsteori og pedagogikk gir mange eksempler innen organisasjonslæring, et felt innen læringsteori og pedagogikk som skiller seg fra pedagogikk og didaktikk i en opplærings- og utdanningskontekst. Utviklingsarbeid i profesjonsfellesskap er betinget av et felles begrepsregister med noenlunde felles forståelse. Det er behov for å utvikle både testmetoder og mekanismer for å validere kvaliteten i responsen til språkmodeller, før responsen fra de tillegges betydningsfull vekt.

- Gi god informasjon om kjente begrensninger med respons gitt av kunstig intelligens, inkludert at særlig generative KI-systemer er designet for å gi variasjon i respons og at denne ikke nødvendigvis er presis.

8.7 LAGRINGSBEGRENSNING

Det er flere aktuelle problemstillinger rundt lagringsbegrensning i KI-applikasjoner. Problemstillingen er tett forbundet med formålsbegrensning, i at personopplysninger og andre data lagret lengre enn nødvendig, øker risikoen for at de behandles for andre formål enn tiltenkt. For eksempel så er det vesentlig å avklare hvorvidt en KI-tjeneste lagrer opplysninger i skytjenesten der modell- og inferensinfrastrukturen befinner seg, eller at inn- og utdata lagres i andre skytjenester, etter at de er ferdig behandlet i modellen og brukeren har mottatt responsen. Det er vanlig praksis for tilbydere av KI-tjenester å lagre inn- og utdata i 30 dager. Også andre logger vil i mange tilfeller inneholde personopplysninger, og kan typisk oppbevares i lengre perioder, enten av sikkerhetshensyn, for å avdekke brudd på tjenestevilkår eller for å imøtekomme lovkrav. Slike data analyseres og benyttes både maskinelt og av mennesker.

Behandlingsansvarlig kan i liten grad påvirke disse lagrings- og sletterutinene, som kan få innvirkning på behandlingsaktivitetene som utføres. Det må vurderes i hvilken grad de anses

¹² <https://www.udir.no/laring-og-trivsel/lareplanverket/forsok-og-pagaende-arbeid/Retningslinjer-for-utforming-av-lareplaner-for-fag-/vedlegg/#verbbeskrivelse>

<https://www.udir.no/laring-og-trivsel/lareplanverket/kjennetegn/kjennetegn-pa-maloppnaelse--matematikk-1p-vg1/>

som vesentlige, og dermed skal defineres av behandlingsansvarlig, eller er av en ikke-vesentlig art, som kan overlates til databehandler å beslutte.

- Det bør vurderes begrenset hvilke kategorier av personopplysninger som behandles i tjenester der det er begrenset tilgang til å påvirke lagrings- og sletterutiner.

8.8 INTEGRITET OG KONFIDENSIALITET

Ved anskaffelse av KI-tjenester er det viktig å ha god oversikt over og kontroll på behandlingsaktiviteter knyttet til input, altså det som sendes f.eks. en språkmodell. Det er også viktig å kun velge databehandlere som det er grunn til å tro at forstår instruks fra behandlingsansvarlig (skoleeier) og sitt ansvar i tråd med artikkel 28 i personvernforordningen. Særlig oppmerksomhet burde rettes mot KI-tjenestens mulighet til innsikt i innhold og metadata sendt til og fra en språkmodell eller annen form for KI, og dermed også potensialet for at opplysninger kan viderebehandles, endres og tilgjengeliggjøres i systemer eller for personer der de ikke skal behandles eller gjøres kjent. Risiko varierer i stor grad med hvilke kapabiliteter et system har, og i hvor stor grad brukerne kan definere funksjonalitet som iverksetter nye behandlingsaktiviteter. KI-tjenester tilbys ikke nødvendigvis som granulerte enkeltkomponenter, men mer overordnede systemer med et bredt sett med funksjonalitet. Tjenesteleverandør forbeholder seg i stor grad retten til å endre de bakenforliggende komponentene, noe som gjør at det kan være en utfordring å vurdere og implementere tiltak som er effektive over tid.

Agentiske KI-tjenester utgjør en særlig risiko for integriteten og konfidensialiteten til personopplysninger i digitale læringsressurser og pedagogiske fagsystem. Slike systemer kan berike utdata gjennom søk i eksterne datakilder (RAG). Agentiske og berikede systemer kan gjøre responsen fra KI mer relevant, korrekt og nyttig, men utgjør også en særlig og økende kilde til risiko for konfidensialitetsbrudd. Agentiske systemer kan utvikle nye kapabiliteter og utføre behandlinger, uten at en bruker har instruert disse til å utføre slike handlinger og uten at det nødvendigvis er synlig for den. Dersom datakilder med personopplysninger er tilknyttet disse, så medfører det en risiko for at de både kan endres eller tilgjengeliggjøres.

- KI-tjenester som tas i bruk bør avvikles fullstendig ved endt bruk.
- Ha kontroll på hvor KI-tjenester kan lese og skrive data, samt hvilke operasjoner/funksjoner systemet kan utføre.

8.9 ANSVARLIGHET OG BEHANDLINGSANSVARLIGES VIDERE VURDERINGER

Å foreta en vurdering av personvernkonsekvenser for bruk av kunstig intelligens i digitale læringsressurser har vært et forsøk i å knytte seg til et objekt i høy fart. Vurderingen har tatt høyde for de behandlingsaktiviteter som er identifisert som mest aktuelle og innenfor rammen av prosjektet. Ettersom kapabilitetene til den underliggende teknologien stadig utvikler seg, så noen av vurderingene kan ha begrenset verdi fremover i tid dersom de ikke oppdateres. Underdatabehandler vil oppdatere vilkår for tjenesten, som databehandler må følge og innrette

seg etter. Det er stor sannsynlighet for at teknologien har utviklet seg vesentlig, allerede innen denne vurderingen tas inn i en implementeringsprosess i en digital læringsressurs eller pedagogisk fagsystem. Det vil nok allerede i inneværende år være elementer i vurderingen som er datert, og som trenger oppdatering.

For implementering av teknologien f.o.m skoleåret 2024/25, så vil det i hovedsak være tilstrekkelig å ivareta behandlingsaktivitetene iht. gjeldende lovkrav, primært i personvernforordningen, med opplæringslova og forvaltningslov med forskrifter. Med tanke på typisk implementeringsperiode i skolen, er det uansett viktig å være klar over lovkrav som potensielt vil inntreffe. KI-forordningen (EU 2024/1689) vil stille nye krav til åpenhet, samt for mange bruksscenarioer i oppvekst, opplæring og utdanning, kreve nye former for vurderinger, utover minstekravene til personvernkonsekvensvurderingen. Særlig for sektoren en forpliktelse til å påse at teknologien ikke gjenskaper eller forsterker diskriminering, gjennom handlinger som bygger på vektning og bias i KI-modellene. (AlgorithmWatch, 2023)

Risikonivået som den behandlingsansvarlige aksepterer vil i stor grad være definert av hvor sentralt KI-integrerte læringsressurser er i opplæringen på den enkelte skole, og dermed hvor stor rolle behandlingene spiller i oppfyllelse av formålet om undervisvurdering. «Undervisvurdering i fag skal være en integrert del av opplæringen, og skal brukes til å fremme læring, tilpasse opplæringen og øke kompetansen i fag.» (Utdanningsdirektoratet, 2024a) Vurdering skal være en integrert del av opplæringen, og prinsippet om undervisvurdering er som en integrert del av opplæringen. En del KI-tjenester brukt i opplæring og utdanning er ikke spesifikt utviklet til bruk i den konteksten, men kan benyttes til å oppfylle formålene. Andre tjenester er utviklet med den hensikt å brukes innen opplæring og utdanning, og har funksjoner spesielt utviklet for pedagogiske formål. Det er ikke vurdert her om de generelle eller spesifikke systemene utgjør høyere eller lavere grad av risiko til de registrerte, men det er sannsynlig at risikoene kan være ulike, både i form og situasjon. Den underliggende teknologien for kunstig intelligens vil i mange tilfeller være svært lik, sett bort fra videretrening og annen optimalisering, slik at det nærliggende å se til implementeringen av teknologien for å identifisere risikoer som er særlige for opplæring og utdanning.


En vesentlig del av risikoen ved bruk av kunstig intelligens i digitale læringsressurser og fagsystem oppstår i den situasjonen der systemet tas i bruk for ulike formål. I opplæring og utdanning så kan det blant annet være i undervisning eller ved en eksamenssensur. Der det er noe enklere for utviklere og for forvaltningsorgan å påse at tilstrekkelig sikkerhet er implementert i overføring og lagring av opplysninger, i tjenesten, så er det vanskeligere å påse at systemet ikke brukes på måter som kan forringe elevenes rettigheter og friheter. Det er vesentlig hvilken opplæring særlig lærer og elever får, og i hvilken retning både den underliggende teknologien og tjenestene «på toppen» utvikler seg, og sektorens mulighet til å tilpasse seg dette. Ansatte og elever som skal ta i bruk, eller være gjenstand for behandlinger i høyrisikosystemer, vil ha krav på særskilt informasjon om systemene og behandlingene. God informasjon og veiledninger er derfor viktige forutsetninger for å utvikle kompetansen som er nødvendig for å få et godt utbytte, brukt på en ansvarlig måte.

9 KILDER

AlgorithmWatch. (2023). 200 students failed their exams. Automated proctoring could be to blame, but doubts remain. <https://algorithmwatch.org/en/spain-students-failed-blame-automated-proctoring/>

Anthropic. (2024). Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. Anthropic. <https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html%2021.5.2024>

Apple. (2024. 24.10.24). Combining Machine Learning and Homomorphic Encryption in the Apple Ecosystem. <https://machinelearning.apple.com/research/homomorphic-encryption>

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, mars). On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (s. 610-623).

Burgess, M. (2023, 23. november). OpenAI's Custom Chatbots Are Leaking Their Secrets. *Wired*. <https://www.wired.com/story/openai-custom-chatbots-gpts-prompt-injection-attacks/>

Carlini et al. (2021). Extracting training data from large language models. <https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting>

Carlini et al. (2023). Quantifying Memorization Across Neural Language Models. <https://arxiv.org/abs/2202.07646>

Castelfranchi, C. (1998). Modelling social action for AI agents. *Artificial Intelligence*. Volume 103, Issues 1–2. <https://www.sciencedirect.com/science/article/pii/S0004370298000563>

Chai et al. (2021). Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning With Applications*. vol 6. <https://www.sciencedirect.com/science/article/pii/S2666827021000670#b11>

Cohen, P. R., & Levesque, H. J. (1995). Communicative Actions for Artificial Agents. In *ICMAS* (Vol. 95, pp. 65-72). <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=e7925577647a39817f58a0c6693564af7ca00482>

Datatilsynet. (2023a). *De 10 brud*, <https://www.datatilsynet.dk/hvad-siger-reglerne/vejledning/sikkerhed/de-10-brud>

Datatilsynet. (2023b). *Offentlige myndigheders brug af kunstig intelligens*. <https://www.datatilsynet.dk/Media/638321084132236143/Offentlige%20myndigheders%20brug%20af%20kunstig%20intelligens%20-%20Inden%20I%20g%C3%A5r%20i%20gang.pdf>

Datatilsynet. (2024a). *Offentliggørelse af datasæt og AI-model*. <https://www.datatilsynet.dk/afgoerelser/afgoerelser/2024/jan/offentliggoerelse-af-datasæt-og-ai-model>

- Datatilsynet. (2024b). Datatilsynet giver påbud i Chromebook-sag. <https://www.datatilsynet.dk/afgoerelser/afgoerelser/2024/jan/datatilsynet-giver-paabud-i-chromebook-sag>
- Digdir. (19.02.2024). Oversikt over EU-regelverk om deling og bruk av data. <https://www.digdir.no/datadeling/oversikt-over-eu-regelverk-om-deling-og-bruk-av-data/3251>
- EDPB. Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is “likely to result in a high risk” for the purposes of Regulation 2016/679, 2017. https://edpb.europa.eu/our-work-tools/our-documents/guidelines/data-protection-impact-assessments-high-risk-processing_en
- Edwards, B. (2024a). AI-generated articles prompt Wikipedia to downgrade CNET’s reliability rating. Ars Technica. <https://arstechnica.com/information-technology/2024/02/wikipedia-downgrades-cnets-reliability-rating-after-ai-generated-articles/>
- Edwards, B. (2024b). Google’s hidden AI diversity prompts lead to outcry over historically inaccurate images. <https://arstechnica.com/information-technology/2024/02/googles-hidden-ai-diversity-prompts-lead-to-outcry-over-historically-inaccurate-images/>
- European Commission. 15.04.2024. [What is personal data? - European Commission \(europa.eu\)](https://european-commission.eu/what-is-personal-data/)
- European Parliament. (2024). Artificial Intelligence Act. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689
- Gage, P. (1994). A new algorithm for data compression. *The C Users Journal*, 12(2), 23-38.
- Gao et al. (2024). Scaling and evaluating sparse autoencoders. 6.6.2024. <https://arxiv.org/html/2406.04093v1>
- Gregor et al. (2014). Deep AutoRegressive Networks. Xing, E. *Proceedings of the 31st International Conference on Machine Learning*, <http://proceedings.mlr.press/v32/gregor14.html>
- GreenbergTraurig. (26.7.2022). Application of the CLOUD Act to EU Entities. <https://english.ncsc.nl/latest/weblog/weblog/2022/how-the-cloud-act-works-in-data-storage-in-europe>
- Greshake et al. (2023). Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. <https://ui.adsabs.harvard.edu/abs/2023arXiv230212173G/abstract>
- Grube. (2015). Sticky words? Towards a theory of rhetorical path dependency. <https://www.tandfonline.com/doi/abs/10.1080/10361146.2016.1171824>
- Harang, R. (2023). Best Practices for Securing LLM-Enabled Applications. <https://developer.nvidia.com/blog/best-practices-for-securing-llm-enabled-applications/>
- Huang et al. (2022). ConceptExplainer: Interactive Explanation for Deep Neural Networks from a Concept Perspective. <https://ieeexplore.ieee.org/abstract/document/9903285>

Huang et al. (2023) Language Is Not All You Need: Aligning Perception with Language Models. A. Oh et al. Advances in Neural Information Processing Systems 36 (NeurIPS 2023)

<https://arxiv.org/abs/2302.14045>

Huang, K. (2023). Mitigating Security Risks in Retrieval Augmented Generation (RAG) LLM Applications. <https://cloudsecurityalliance.org/blog/2023/11/22/mitigating-security-risks-in-retrieval-augmented-generation-rag-llm-applications>

Kai et al. (2023). Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection.

<https://ui.adsabs.harvard.edu/abs/2023arXiv230212173G/abstract>

Kelleher, John D. (2019). Deep Learning. The MIT Press.

Kharitonov, E., Baroni, M., & Hupkes, D. (2021). How bpe affects memorization in transformers. arXiv preprint arXiv:2110.02782.

Kuhn et al. (2023). Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. International Conference on Learning Representations 2023.

<https://arxiv.org/abs/2302.09664>

Lewis et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.

<https://arxiv.org/abs/2005.11401v4>

Longpre et al. (2024). A Safe Harbor for AI Evaluation and Red Teaming. Preprint.

<https://sites.mit.edu/ai-safe-harbor/>

Medietilsynet. (2021, 19 mai). Dette er kritisk medieforståelse.

<https://www.medietilsynet.no/digitale-medier/kritisk-medieforstaelse/dette-er-kritisk-medieforstaelse/>

Mirzadeh et al. (2024). GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models

Ng, A. (2011). Sparse autoencoder. <https://graphics.stanford.edu/courses/cs233-21-spring/ReferencedPapers/SAE.pdf>

NOU 2023:19. (2023). *Læring, hvor ble det av deg i alt mylderet?* Kunnskapsdepartementet

Park et al. (2023). Generative Agents: Interactive Simulacra of Human Behavior. UIST '23: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology.

<https://dl.acm.org/doi/abs/10.1145/3586183.3606763>

Paterson, A. (2024, 11. januar). Wolf in Sheep's Clothing? Security Implications of OpenAI's GPT Store. <https://www.harmonic.security/post/wolf-in-sheeps-clothing-security-implications-of-openais-gpt-store>

Radford, A. (2018). Improving language understanding by generative pre-training. Preprint.

Ramjee et al. (2024). CataractBot: An LLM-Powered Expert-in-the-Loop Chatbot for Cataract Patients. arXiv:2402.04620 <https://arxiv.org/abs/2402.04620>

Regjeringen. (2024, 29. januar). Forslag til forordning om kunstig intelligens (KI-forordningen). <https://www.regjeringen.no/no/sub/eos-notatbasen/notatene/2021/juni/forslag-til-forordning-om-kunstig-intelligens-ki-forordningen/id2884935/>

Russel and Norvig. (2010). Artificial Intelligence A Modern Approach *Third Edition*. Pearson Education

Schaul et al. 2023. Inside the secret list of websites that make AI like ChatGPT sound smart <https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/>

Schroeder, P. 6.7.2023. Multimodal Augmented Inputs in LLMs using Azure Cognitive Services. Microsoft. [Multimodal Augmented Inputs in LLMs using Azure Cognitive Services - Surface Duo Blog \(microsoft.com\)](https://www.microsoft.com/en-us/blog/2023/07/06/multimodal-augmented-inputs-in-llms-using-azure-cognitive-services-surface-duo/)

SLM Rijk. (2024). DPIA Microsoft 365 Copilot. <https://slmmicrosoftrijk.nl/wp-content/uploads/2024/12/20241218-Public-version-DPIA-Microsoft-365-Copilot-for-SLM-Rijk.pdf>

Staab et al. (2023). Beyond Memorization: Violating Privacy Via Inference with Large Language Models. <https://arxiv.org/abs/2310.07298>

Thakkar, O. D., Ramaswamy, S., Mathews, R., & Beaufays, F. (2021, Juni). Understanding unintended memorization in language models under federated learning. In Proceedings of the Third Workshop on Privacy in Natural Language Processing (s. 1-10).

Utdanningsdirektoratet. (2024a). Underveisvurdering. <https://www.udir.no/laring-og-trivsel/vurdering/om-vurdering/underveisvurdering/>

Utdanningsdirektoratet. (2024b). Rammeverk for lærerens profesjonsfaglige digitale kompetanse. <https://www.udir.no/kvalitet-og-kompetanse/digitalisering/rammeverk-larerens-profesjonsfaglige-digitale-komp/>

Vaswani et al. (2017). Attention is all you need. I. Guyon et al. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. <https://arxiv.org/abs/1706.03762>

Vincent, J. (2023). Meta's powerful AI language model has leaked online — what happens now?. The Verge. <https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse>

Wang et al. (2023). Voyager: An Open-Ended Embodied Agent with Large Language Models. Preprint. <https://arxiv.org/abs/2305.16291>

White, J. (2023). How Strangers Got My Email Address From ChatGPT's Model. New York Times. <https://www.nytimes.com/interactive/2023/12/22/technology/openai-chatgpt-privacy-exploit.html>

Wikipedia. 2024. Reliable sources/Perennial sources. https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources/Perennial_sources

Yang et al. (2023). Diffusion Models: A Comprehensive Survey of Methods and Applications. ACM Computing Surveys. Volume 56. <https://dl.acm.org/doi/abs/10.1145/3626235>

10 VEDLEGG

Vedlegg 1: KI-ordliste

Autoregressiv modell: En prediksjonsmodell som vil beregne med en grad av sannsynlighet hva som skal komme, basert på verdier som representerer det som har vært.

Konfabulasjoner: Ofte også betegnet som «hallusinasjoner» i omtale av KI som generer en respons som vi oppfatter av burde være faktamessig korrekt, men som kun er et resultat av at modellen har beregnet en viss token som mest sannsynlig.

Kunstig intelligens: En samlebetegnelse for resultatet av maskinlæring eller kompleks informasjonsbehandling.

Språkmodell: en maskinlæringsmodell som er trent opp til å predikere/forutsi (det mest sannsynlige) neste ord gitt de foregående ordene.

Multimodal: En maskinlæringsmodell som kan behandle flere modaliteter av informasjon, f.eks. skrift, lyd, bilde, video.

Samtalerobot: Brukerrettet tjeneste eller programvare med et grensesnitt for dialog med en eller flere språkmodeller. Teknologi som er dialogisk av natur.

Øvrige begrep se:

Datatilsynets ordliste: <https://www.datatilsynet.no/regelverk-og-verktoy/ordliste>

Teknologirådets ordliste for kunstig intelligens: <https://teknologiradet.no/ordliste-for-kunstig-intelligens/>