De nasjonale
FORSKNINGSETISKE
KOMITEENE

Forskningsetisk betenkning om kunstig intelligens

NENT • Den nasjonale forskningsetiske komité for naturvitenskap og teknologi

Forskningsetisk betenkning om kunstig intelligens
1. utgave – november 2019

ISBN: 978-82-7682-097-3

Forsidebilde:
Et ubemannet helikopter klargjøres for oppdrag på det amerikanske krigs- skipet USS Coronado. Foto: Flickr/COMSEVENTHFLT, lisens CC BY-SA 2.0

FORSKNINGSETISK BETENKNING OM KUNSTIG INTELLIGENS

Den nasjonale forskningsetiske komité for naturvitenskap og teknologi (NENT) avgir med dette en forskningsetisk betenkning om kunstig intelligens (KI).

Det er et uttalt mål for samtlige stormakter å bli ledende innen KI, og dette forskningsfeltet er i rask utvikling. KI påvirker allerede i dag de fleste områder av samfunnet. Samtidig har fremtredende forskere og næringslivsledere uttrykt bekymring for utviklingen, spesielt når det gjelder selvlærende systemer som ikke bare erstatter rutinepregede handlinger, men radikalt endrer og utvider det menneskelige handlingsrom. Den videre utviklingen av KI-teknologi og den usikkerheten som er knyttet til de konsekvensene den vil ha for mennesker og samfunn, krever en refleksjon som også forskere må forholde seg til. Dette danner bakgrunnen for NENTs betenkning.

Oslo, november 2019

NENT (2018-2021)
Øyvind Mikkelsen (komitéleder), Michaela Aschan, Ingrid Bay-Larsen, Tone Druglitrø, Ole Andreas Engen, Hanne Pernille Guldbrandsen, Steinar Heldal, Kjellrun Hiis Hauge, Gorm Idar Johansen, Cecile Marie Mejdell, Rune Nydal, Jørn Paus, Ketil Skogen, Jim Tørresen, Lise Øvreås, samt Helene Ingierd (sekretariatsleder).

Innledning

Denne betenkningen peker ut og beskriver de spesielle utfordringene som gjør seg gjeldende i forbindelse med forskning om KI, og hvilke forskningsetiske spørsmål disse gir opphav til. *Forskningsetiske retningslinjer for naturvitenskap og teknologi* vektlegger forskningens selvstendige ansvar for den rollen den har i samfunnsutviklingen, og betenkningen utdyper spesielt hvordan forskningens samfunnsansvar bør forstås i lys av de utfordringene KI-forskning reiser.

I henhold til Lov om organisering av forskningsetisk arbeid (forskningsetikkloven) har forskere og forskningsinstitusjoner et selvstendig ansvar for å sikre at den forskningen de utfører er etisk forsvarlig. NENT er et faglig uavhengig organ som skal gi råd til forskere og myndigheter om forskningsetiske spørsmål. Ambisjonen med denne betenkningen er mer spesifikt å tilrettelegge for god og ansvarlig KI-forskning i Norge. Betenkningen er primært skrevet for forskere, forskningsinstitusjoner og andre aktører som legger premisser for eller er involvert i KI-forskning. Betenkningen bør ses i sammenheng med andre deler av De nasjonale forskningsetiske komiteenes arbeid for å identifisere forskningsetiske utfordringer ved digitalisering og stordata (NESH, Forskningsetisk veileder for internettforskning 2019; FEK, Rapport om stordata, *kommer* 2020).

Sammendrag

1. Sikre menneskeverd
Forskere og forskningsinstitusjoner må sikre at KI-systemene er bygd opp på en måte som ivaretar individenes selvbestemmelse, menneskeverd og demokratiske rettigheter. Likeledes må forskere vurdere forventet eller mulig innvirkning på enkeltpersoner, dyr, miljø og samfunn, og det må legges til rette for rettferdig og etisk bruk av systemene.

2. Lokalisere ansvar
Forskere som utvikler og designer KI-systemer kan legge føringer på de beslutningene systemene tar handlingene de utfører, og fagmiljøene har derfor et spesielt ansvar. Ved oppdragsforskning eller planlagt kommersialisering av forskningsresultater bør forskere samarbeide med eksterne aktører for å vurdere risiko ved videre bruk av forskningen.

3. Inspiserbarhet
Inspiserbarhet, dvs. muligheten til å identifisere kildene til de dataene som brukes og genereres av systemene, samt hvordan systemene tar beslutninger, er avgjørende for å sikre hensyn til rettferdighet og tillit når avgjørelser tas automatisk. Forskere bør synliggjøre og begrunne manglende inspiserbarhet, og KI-forskningen bør ha som mål å frembringe «glassbokser», dvs. systemer som lar seg inspisere.

4. Forskningsformidling
Det er utfordrende å sikre en balansert diskusjon om de risikoene og mulighetene KI innebærer. Som samfunn bør vi unngå naivitet, og samtidig være klar over mulige risikoer og muligheter, for eksempel for at KI kan komme i feil hender. Forskere skal bidra til en informert samfunnsdebatt, slik at samfunnets vurderinger kan baseres på realistiske forutsetninger. Forskere har et særlig ansvar for å formidle risikoer og muligheter på en balansert måte, siden de har best kunnskap om hvor langt utviklingen har kommet.

5. Erkjenne usikkerhet
Forskningsetisk sett er det vesentlig å vurdere og kommunisere den usikkerheten som er forbundet med forskning. Utviklingen av KI er

kjennetegnet av grunnleggende usikkerhet og uforutsigbarhet. NENT ser derfor et behov for systematiske studier av de risikoene som er forbundet med utviklingen av KI. Myndigheter og andre som finansierer forskning, bør legge til rette for tverrfaglighet i forskningen og derved erkjenne dens uforutsigbarhet og minimere usikkerhet der dette er mulig. Etikk bør tas inn som et fag i utdanningen av fremtidige utviklere av KI.

6. Sikre bred involvering
Forskere har også et ansvar for å kommunisere den risikoen som følger av deres forskningsfunn. Hvilke risikoer og muligheter ved teknologien som vektlegges, kan også avhenge av hvilket etisk perspektiv og hvilke verdier og interesser som legges til grunn. De som blir mest påvirket av de beslutningene som tas, må sikres en stemme i beslutningsprosessene. Myndigheter og forskningsinstitusjoner bør legge til rette for bred involvering av innbyggerne i diskusjoner om hva som er formålet med forskningen, innretningen på forskningsprogrammene og bruken av forskningen.

7. Sikre personvern og hensynet til enkeltmennesker
Grunnleggende personvernprinsipper, nedfelt i personvernlovgivningen, skal følges. Forskningsetisk sett er samtykke en hovedregel når personopplysninger brukes i forskning. Selv om det benyttes anonymiserte data i analysene vil sammenstillinger med andre data likevel kunne avdekke sensitive opplysninger eller avsløre enkeltpersoner, og dermed utgjøre personlige data. Innhenting og bruk av data som inkluderer persondata kan utfordre kravet om informert samtykke. Ved innsamling og nye sammenstillinger av store datamengder er det en særlig risiko for at personlig informasjon kan brukes på måter vi ikke kjenner til (fordi formålet også er ukjent for forskeren på tidspunktet for innsamlingen), og som vi kanskje ikke ønsker.

I sine forskningsetiske vurderinger knyttet til informasjon og samtykke har forskere et ansvar for å vurdere opplysningenes grad av offentlighet, informasjonens sensitivitet, de berørtes sårbarhet og forskningens interaksjon og konsekvenser (NESH 2019).

8. Kvalitetssikring
I KI-forskning kan det være særlig grunn til å stille kritiske spørsmål ved dataenes kvalitet, sannferdighet og relevans, fordi vi ikke alltid kjenner kildene til dataene, og metadata kan mangle eller være usikre. Skjevheter i materialet, egenskaper ved analyseverktøyet og menneskelige fortolkninger øker mulighetene for feilslutninger. Dette gir grunnlag for usikkerhet i fortolkninger og beslutninger som er basert på KI. For å sikre etterprøvbarhet og kvalitet bør forskere og forskningsinstitusjoner derfor legge til rette for at datakilder skal være åpne og allment tilgjengelige.

9. Rettferdig tilgang til data
Forskningsetisk sett er det vesentlig å legge til rette for at forskningen, inkludert data og resultater, som hovedregel gjøres tilgjengelig for alle. NENT ser en risiko for at store deler av forskningsinnsatsen innen KI unndrar seg de kravene til åpenhet som gjelder for forskning ellers (slik de er nedfelt i bl.a. FAIR-prinsippene), for eksempel under henvisning til behovet for hemmelighold av konkurransefortrinn. Myndigheter og forskningsinstitusjoner bør legge til rette for allmenn tilgang til data. De bør sørge for åpenhet om hvem som skal ha eierskap til teknologi, infrastruktur og data, hvilke forskningsområder som blir prioritert og hvorfor, og hvem som kan forventes å ha nytte av forskningsinnsatsen.

Betenkningen har følgende struktur: Etter en kort redegjørelse for den metoden NENT har lagt til grunn, utdyper vi hva som kjennetegner KI-forskningen i dag. Deretter spør vi hvilke utfordringer KI representerer i samfunnet og forskningshverdagen, og hvilke forskningsetiske implikasjoner disse fører med seg.

Metode

NENT har vært i dialog med aktuelle fagmiljøer for KI-forskning gjennom en høringsrunde juni-august 2018 og en workshop i februar 2019. Formålet var å kartlegge hva de norske forskningsmiljøene ser som de sentrale mulighetene og forskningsetiske utfordringene.
NENT ba om følgende innspill:
1. Utfører dere forskning som dere vil si ligger innenfor området KI, og hvilke forskningsmiljøer er engasjert i dette?
2. Hva anser dere som positive muligheter for KI? Ser dere noen bekymringsfulle sider ved utviklingen av KI?
3. Hvilke forskningsetiske spørsmål og utfordringer (inkludert spørsmål om samfunnsmessige konsekvenser) gjør seg gjeldende i KI-forskning, inkludert deres egen?
4. Hva bør forskere og forskningsinstitusjoner selv ta ansvar for i en bærekraftig utvikling av feltet, dvs. en utvikling som fremmer nytenkning og kunnskap, og samtidig sikrer at forskningsetiske hensyn blir ivaretatt?

NENT mottok 13 innspill, og sammen med den workshopen som komiteen arrangerte, har dette utgjort et viktig bakteppe for kartleggingen av hva forskningsmiljøene selv anser som sentrale utfordringer ved KI. I tillegg har NENT orientert seg om hva som er gjort på dette området nasjonalt og internasjonalt. Noen av de mest sentrale dokumentene som er utarbeidet internasjonalt pr. i dag, og som NENT har gått gjennom, inkluderer:

- The Asilomar AI Principles, The Future of Life Institute, 2017
- The General Principles in *Ethically Aligned Design* (V2), IEEE, 2017
- Report on Robotics Ethics, COMEST, 2017
- Towards a Digital Ethics, EDPS, 2017
- Code of Ethics and Professional Conduct, ACM (2018)
- The Ethical Principles in *Statement on Artificial Intelligence*, EGE, 2018
- Guidelines for trustworthy AI, High-Level Expert Group on AI, European Commission, 2019.
- Principles on AI, OECD, 2019

Dokumentene har til dels ulikt nedslagsfelt; enkelte retter seg mot området KI som helhet, mens andre behandler tilstøtende og til dels overlappende områder, som autonome og intelligente systemer (IEEE), robotikk (COMEST) og digital teknologi generelt (EDPS). Så godt som samtlige av dokumentene over har vært gjenstand for brede innspillsrunder med deltakere fra forskersamfunnet, industrien, politiske organer og andre interessenter («stakeholders»). I næringslivet har selskaper som IBM, Microsoft og Googles Deep Mind utviklet egne etiske retningslinjer, og gått sammen om å utvikle brede initiativer som «Partnership on AI» og «OpenAI».

I Norge har Datatilsynet utgitt en rapport om KI og personvern, mens Teknologirådet har lagt frem rapporten «KI – muligheter, utfordringer og en plan for Norge» som også tar for seg etiske aspekter av KI (begge utgitt i 2018). Regjeringen har bestemt at det skal utvikles en nasjonal strategi for KI, der etiske perspektiver også vil stå sentralt. Dette er i tråd med utviklingen i de fleste europeiske land, der slike strategier allerede er utviklet eller vil bli utviklet innen utgangen av 2020.

NENT mener det er viktig å merke seg det økende antallet rapporter om etikk og KI, noe som tyder på at det finnes en sterk bevissthet om behovet for etisk refleksjon rundt denne teknologien. De forskningsetiske aspektene knyttet til KI er derimot lite utviklet, og NENT ser et behov for å utdype disse implikasjonene av teknologiutviklingen. I NENTs vurderinger av KI-forskning har forskningsetiske retningslinjer, spesielt *Forskningsetiske retningslinjer for naturvitenskap og teknologi*, utgjort et rammeverk. Etiske vurderinger knyttet til KI kan dels handle om spørsmål som oppstår i selve utviklingen av teknologien, dels om spørsmål som oppstår ved den videre bruken av den. Slik den er definert i retningslinjene omfatter forskningsetikken også det sistnevnte spørsmålet, dvs. den fordrer en etisk refleksjon utover selve forskningsprosessen.

Kjennetegn ved KI

KI har eksistert som forskningsfelt i rundt 60 år innenfor datateknologi og informatikk, og har som målsetting å virkeliggjøre KI. Bredt definert omfatter

KI teknikker som er utviklet for å la datamaskiner inngå i teknologiske systemer på måter som gjør at de oppfører seg «intelligent», dvs. at de er i stand til å løse kognitive og fysiske oppgaver som tidligere har vært forbeholdt mennesker. Dette skjer gjennom dataprogrammer som benytter data og algoritmer til å trene opp eller optimalisere et system til å gi en ønsket respons – enten i en utviklingsfase eller etter at systemet er tatt i bruk. Et eksempel kan være talegjenkjenning på mobiltelefoner; systemet blir stadig flinkere etter hvert som brukeren korrigerer feil.

Et helt sentralt og særegent kjennetegn ved KI er dermed at slike systemer *etterligner, erstatter og utvider menneskelig intelligent handling, og menneskelig beslutningstaking og vurdering*. Teknologien har også potensial til å identifisere og simulere menneskelig følelser på måter som gjør at vi opplever at maskinen har menneskelige trekk. I den ene enden av skalaen finner vi «*deterministiske systemer*» som erstatter rutinepregede handlinger, og i den andre enden har vi «kognitive» eller «komplekse og helt autonome» systemer som erstatter handlinger vi forbinder med menneskelig vurdering, resonnering og læring (COMEST, 2017, s.7). Dette skillet innbefatter også ulike grader av automatiserte systemer, dvs. systemer som kan operere «selvstendig» eller «autonomt», uten menneskelig inngripen. Denne skalaen strekker seg fra systemer som blir fjernstyrt av en operatør, til helt autonome systemer som tar alle beslutninger selv, ut fra et oppdrag de har fått av en operatør. Den første typen av systemer vil i stor grad kunne være forhåndsprogrammerte, mens den andre gruppen systemer vil som oftest ha behov for å fortsette å lære mens de er i bruk. Graden av kompleksitet angår både hva et systemet må være i stand til å kunne oppfatte og utføre. Et annet beslektet skille går mellom spesifikk og generell KI. På den ene siden kan KI utføre ganske enkle tjenester, slik som å tilpasse anbefalinger gitt av ulike nettjenester, spille sjakk eller gjenkjenne ansikter. På den andre siden vil det imidlertid kunne finnes systemer som er i stand til å utføre mange forskjellige oppgaver. Generell KI i sin mest utviklede form omtales gjerne som superintelligens, ofte forbundet med en forestilling om fullt utviklede «bevisste» maskiner. Mens utviklingen innen spesifikk KI har gått raskt de siste årene, har det vært gjort få fremskritt innen generell KI. Hvorvidt vi i det hele tatt vil kunne utvikle generell KI og superintelligens, og i så fall når det vil skje, er et omstridt spørsmål.

Et annet viktig kjennetegn er at KI har *utallige anvendelsesområder og et enormt endringspotensial*. Utviklingen skjer dels i samhandling med mennesker, og dels i interaksjon mellom de tekniske systemene. Teknologien er dermed også kjennetegnet av *uforutsigbarhet*; det er vanskelig eller umulig å forutse hvilke virkninger teknologien vil ha for enkeltindivider, samfunnet og miljøet. I mange land er implementeringen av teknologien allerede godt i gang på områder som helsevesen, rettsvesen, transport og kommunikasjon, mens andre områder er under utvikling. For eksempel på området helse kan KI i dag vurdere bilder av mulig føflekkreft bedre enn det mange hudleger er i stand til. Ved Universitetet i Agder og Universitetet i Oslo forskes det på hvordan KI kan brukes i psykologisk helsehjelp. Det antas at det finnes utallige bruksområder knyttet til helse, for eksempel ved komplekse kirurgiske inngrep og i omsorg for pasienter. Det forventes at KI på liknende måter vil få store konsekvenser også for andre sektorer, som arbeidsmarkedet, økonomien, politikk og kultur. KI vil også kunne påvirke oss som enkeltpersoner, for eksempel slik vi har sett i forbindelse med mobiltelefonenes betydning for mellommenneskelige relasjoner og for hvordan vi tenker og føler.

De senere års utvikling av spesifikk intelligens viser til et tredje kjennetegn ved KI, nemlig *generering av stordata* som kan inneholde personopplysninger. KI i form av maskinlæring (spesielt «dyp læring») er drevet frem av stordata og datakraft. Samtidig er KI en kilde til nye stordata.

Forskningsetiske utfordringer

Den følgende gjennomgangen av ni forskningsetiske utfordringer kan struktureres i tre bolker som gjenspeiler de kjennetegnene ved KI vi har beskrevet ovenfor. Ulike forskningsfelt og anvendelser av KI reiser selvsagt ulike utfordringer, og de forskningsetiske normene som drøftes, vil kanskje være under sterkt press innen enkelte områder, men ikke bli nevneverdig berørt i andre. Punktene er en oppsummering av hva NENT i samarbeid med fagmiljøene har funnet spesiell grunn å rette oppmerksomheten mot i forbindelse med KI-forskning, og de må utdypes med supplerende betraktninger for hvert enkelt forskningsprosjekt.

A) Ansvar for utviklingen og bruken av autonome systemer
Det første settet med vurderinger er knyttet til den målsettingen KI-teknologien har om å etterligne, erstatte og utvide menneskelig intelligent handling og menneskelige beslutninger og vurderinger.

1. Sikre menneskeverd
Menneskeverd innebærer at mennesket skal forstås som mål i seg selv, aldri bare som et middel. Dette setter grenser for definisjoner og kategoriseringer av mennesker på basis av algoritmer og autonome systemer. Utviklingen og bruken av KI berører menneskeverdet på grunnleggende vis; for eksempel gjennom utvikling av smarte hjelpemidler i hverdagslivet kan KI på den ene siden bidra til å fremme enkeltindividers selvrealisering og menneskeverd. På den annen side kan KI også true menneskeverdet. Et eksempel er bruken av KI i overvåkning innenfor et system for sosial kontroll og sanksjonering av befolkningen, slik som det som benyttes i Kina. Utviklingen av algoritmer til bruk i «profilering», dvs. teknikker som benyttes for å analysere, forutsi og eventuelt påvirke fremtidige preferanser og atferdsmønstre, er et annet eksempel. Når individer ikke behandles som mål i seg selv, men som aggregater av data som er samlet for eksempel for å optimalisere den administrative samhandlingen med dem, kan det stilles spørsmål ved om dette lar seg forene med respekt for menneskeverdet (EDPS 2017, s. 16–17). I de senere år har vi også sett flere eksempler på at KI-systemer er blitt brukt til å manipulere demokratiske prosesser, slik som presidentvalget i USA i 2016 og Brexit. Det er viktig at bruken av KI ikke blir en trussel mot demokratiske rettigheter (EGE 2018, s. 17–18).

KI-forskning kan utvikles eller brukes for å fremme enkeltpersoners selvbestemmelse, menneskeverd og demokratiske rettigheter, men gjennom sitt valg av temaer, i forholdet til eventuelle forskningsdeltakere og i formidlingen og bruken av resultater kan den også true disse verdiene. I forskningsetikken er det formulert krav om å hindre og ikke bli delaktig i misbruk av forskning:

> Der vitenskapelig og teknologisk utvikling kan misbrukes til å svekke individenes selvbestemmelse, menneskeverd og demokratiske rettigheter, skal forskeren bestrebe seg på å hindre og ikke bli delaktig i slik misbruk

av forskningen. Forskeren har et selvstendig ansvar for at forskningen direkte eller indirekte vil kunne komme samfunnet til gode og for å minimere risiko (retningslinje 1).

Dette fordrer gode rutiner som kan sikre at det gjøres forskningsetiske vurderinger helt fra starten av forskningsprosessen. «Innebygd etikk» (*Ethics by design*) er et begrep som viser til behovet for en proaktiv tilnærming for å sikre god og ansvarlig KI-forskning. Det bygger på det mer veletablerte begrepet «innebygd personvern», som vi blant annet finner i personvernlovgivningen. «Innebygd etikk» er bredere, og viser til at KI-systemene må være bygd opp på en måte som ivaretar menneskeverd og personvern, inkludert forventede eller mulige virkninger for enkeltpersoner og samfunnet, og at det legges til rette for en rettferdig og etisk bruk av slike systemer. Innebygd etikk innebærer dermed også en vurdering av konteksten til det systemet som utvikles.

2. Lokalisere ansvar

Spørsmål knyttet til muligheten for menneskelig kontroll og plassering av ansvar er grunnleggende i sammenheng med KI, og slike spørsmål er enda mer aktuelle ved utvikling og bruk av adaptive og autonome systemer. Generelt kan det hevdes at jo mer adaptivt og autonomt et KI-system er, jo vanskelige vil det være å kontrollere det, og ansvaret vil bli vanskeligere å lokalisere.

Da en Boeing 737 styrtet i Etiopia i 2018 og 157 mennesker mistet livet, var det fordi pilotene om bord ikke lyktes i å overstyre en alvorlig feil i programvaren MCAS, et antioppbremsingssystem. Denne flyulykken reiser spørsmål knyttet til hvorvidt menneskelig kontroll over systemene var mulig, hvem som har ansvaret når beslutninger er automatiserte, og om et dataprogram kan være ansvarlig for en ulykke.

Når bevegelsene til en maskin styres av et dataprogram, vil ansvaret for utfallene kanskje være mulig å lokalisere. Når det gjelder autonome systemer med dyp læringskapasitet, kan atferds- eller beslutningsprosesser imidlertid ikke programmeres på samme måte som for deterministiske systemer. I debattene om autonome våpensystemer og selvkjørende biler har begrepet

«meningsfull menneskelig kontroll» stått sentralt i spørsmålet om hvem som er ansvarlig. Prinsippet formuleres som en forutsetning for legitimitet, og impliserer at det er mennesker, ikke maskiner eller deres algoritmer, som i siste instans må ha kontroll og stå moralsk ansvarlig. «Meningsfull» refererer gjerne til om et menneske vil ha tilstrekkelig tid til å gripe inn og overstyre maskinen. «Menneskelig kontroll» kan i streng forstand bety at en menneskelig operatør overvåker systemet og tar alle kritiske avgjørelser. I en svakere betydning viser dette til at systemet er utformet slik at det fungerer pålitelig og forutsigbart, uten at et menneske er involvert i hver enkelt beslutning.

Mer spesifikt oppstår det et forskningsetisk spørsmål om hva forskere kan og bør ha kontroll over og ta ansvar for i utviklingen og bruken av adaptive og autonome systemer. Det må skilles mellom ansvaret for KI-forskning på den ene siden og ansvaret for videre bruk av forskningsresultatet på den andre. Forskere som utvikler og designer mer eller mindre autonome KI-systemer, kan legge føringer på de beslutningene systemene tar og de handlingene de utfører. Fagmiljøene har derfor et særlig ansvar.
Forskere har også et medansvar for bruken av forskningen (jf. retningslinje 1-3, 8-9). I forbindelse med oppdragsforskning eller planlagt kommersialisering av forskningsresultater bør forskere derfor samarbeide med eksterne aktører for å vurdere risikoene ved den videre bruken av forskningen.

3. Inspiserbarhet
Betegnelsen «sort boks-problemet» refererer til de ulike utfordringene som er knyttet til at KI-systemene og algoritmene kan være så kompliserte at vi ikke forstår hvordan de har kommet frem til det svaret de gir. «Sort boks-problemet» innebærer dermed en mangel på åpenhet når det gjelder en vesentlig komponent i beslutningsprosessen; vi kjenner kanskje til de dataene som er lagt inn og vi kjenner til svaret, men vi kan ikke avgjøre hvordan dataene har gitt opphav til svaret. Fremveksten av stordata forsterker dette problemet; den mengden av data som legges til grunn, kan være så enorm at vi ikke har mulighet til å få oversikt over den.

Sort boks-problematikken kan også handle om manglende åpenhet om de

betingelsene og rammevilkårene maskinene arbeider ut fra. I utviklingen av autonome systemer brukes matematiske teknikker for å overlate verdivalg til algoritmer. *Deep Mind* har for eksempel utviklet algoritmer ved å bygge på grunnleggende antakelser fra teorien om rasjonelle valg (*rational choice theory*), som ikke uten videre er generelt akseptert.

I innspillene til NENT var *åpenhet* et prinsipp som nær sagt alle trakk frem i sine formuleringer av forskningsetiske utfordringer ved KI.

> Many of the algorithms used in AI are poorly understood, and the applications of AI appear as products of a «black box». As we do not fully understand these algorithms, the applications of AI solutions may lead to unforeseen side effects. These side effects may potentially be dangerous, e.g., if AI based networks is used for medical decisions (Simula).

Mangelen på åpenhet kan gi opphav til at det fattes diskriminerende beslutninger og at visse verdier og perspektiver utelates uten at vi er oppmerksom på det, noe som i sin tur kan innebære manglende tillit til de beslutningene som tas. I denne sammenhengen gir det mening å skille mellom to typer av «sorte bokser»: *For det første* kan dette være en *ufrivillig* sort boks, der mangelen på åpenhet skyldes at modellen er av en slik natur at den ikke lar seg inspisere. *For det andre* kan det være snakk om frivillig skjerming ut fra sikkerhetshensyn, eller at kommersielle aktører ikke ser seg tjent med å offentliggjøre hvilke algoritmer de bruker. I begge disse tilfellene av «sort boks» er problemet at maskinene skjuler hvorfor de tar de valgene de gjør, og at de ansvarlige heller ikke kan forklare bakgrunnen for disse beslutningene. Dette er særlig problematisk når algoritmene i økende grad tar valg som har konsekvenser for enkeltmennesker og samfunnet, for eksempel innenfor rettsvesenet, finanssektoren eller utdanningssektoren. Den norske Skatteetaten benytter for eksempel prediktiv analyse for å velge ut hvilke selvangivelser som bør kontrolleres for mulig juks.

I tillegg til åpenhet blir også andre, dels overlappende prinsipper som transparens eller forklarbarhet også gjennomgående vektlagt som hovedanliggender for en ansvarlig utvikling av KI. Forskningsetisk sett innebærer åpenhet blant annet å være åpen og eksplisitt om valg av

datakilder, utviklingsprosesser og interessenter. *Inspiserbarhet* betegner her mer spesifikt evnen til å beskrive hvordan beslutninger tas av systemene, samt opprinnelsen til de dataene som brukes og genereres av systemet. I henhold til personopplysningslovgivningen er inspiserbarhet også helt avgjørende for å sikre åpenhet og tillit ved automatiserte avgjørelser, som for eksempel «profilering», der maskiner automatisk analyserer eller forutsier forhold ved enkeltpersoner eller grupper (for eksempel knyttet til forhold som økonomi, helse og atferd).

Et system som omfatter en «ufrivillig sort boks» vil ofte kunne gi bedre ytelse enn et mer inspiserbart/transparent system, noe som vil medføre en avveining mellom kvalitet og åpenhet. Det er imidlertid ikke nødvendigvis et spørsmål om å måtte velge det ene fremfor det andre. Forskere bør i alle tilfeller synliggjøre og begrunne slike avveininger, og KI-forskningen bør ha som mål å frembringe «glassbokser», dvs. systemer som lar seg inspisere. Samtidig som det er ønskelig med åpenhet om de bevegelser en maskin gjør, mener NENT forskere har et ansvar for å redegjøre for de antakelsene, valgene og usikkerhetsmomentene som er knyttet til et system (jf. pkt. 5 under).

4. Forskningsformidling

Forsøkene på å identifisere langsiktige konsekvenser av KI-forskningen og bruken av den er beheftet med stor usikkerhet, og kan derfor synes spekulative. I innspillene til NENT understreker flere at de fremstillingene som gjøres av KI i offentligheten gjerne er dystopiske. Problemstillingene oppfattes som konstruerte, da de gjerne er knyttet til utviklingen av generell KI og fullt autonome systemer, mens dagens utvikling i all hovedsak dreier seg om spesifikk KI. På den ene siden kan det hevdes at sannsynligheten for at de mest pessimistiske spådommene slår til er lav, og at hovedfokuset derfor bør rettes mot refleksjon over mulighetene og ulempene ved eksisterende verktøy og data. På den annen side er de mulige skadene knyttet til langsiktige konsekvenser svært store, og dette taler for at vi bør ha et langsiktig perspektiv. Den risikoen som ofte omtales i forbindelse med KI, består i «singularitet», som viser til det punktet i sivilisasjonsutviklingen der KI når et menneskelig nivå av forståelse, og ikke lenger er avhengig av menneskelig interaksjon. Google-sjef Ray Kurzweil har hevdet at vi vil nå

dette punktet innen 2029. Han mener også at med dagens teknologi er vi i ferd med å se begynnelsen av dette punktet. Han viser til den avhengigheten vi er i ferd med å utvikle til telefonene våre, der det neste steget vil bestå i å forbinde teknologien direkte med hjernen (https://www.innomag.no/5-spadommer-fra-googles-fremtidsforsker-ray-kurzweil/). Denne oppfatningen deles av flere fremtredende aktører på feltet. Samtidig er det uenighet, også blant forskere, om generell KI på et menneskelig nivå overhodet er mulig, og i så fall *når*.

Forskere skal bidra til informert samfunnsdebatt, slik at samfunnets vurderinger kan baseres på realistiske forutsetninger. Det er imidlertid utfordrende å oppnå en balansert diskusjon om risikoene og mulighetene ved KI. Fremstillingene av KI i offentligheten kan noen ganger ha karakter av en slags «moralsk panikk» som fremhever scenarier knyttet til muligheten for super- intelligens. På den annen side kan disse mulighetene overdrives, samtidig som de faremomentene som følger med teknologien, kan bli under- kommunisert av de som søker finansiering av utvikling og forskning. Som samfunn bør vi unngå naivitet og være klar over mulige risikoer og muligheter, for eksempel for at KI kan falle i feil hender. Forskere har et særlig ansvar for å formidle risikoer og muligheter på en balansert måte, siden de har best kunnskap om hvor langt utviklingen har kommet.

B) Samfunnsovergripende konsekvenser og forskningens samfunnsansvar
Det andre settet med utfordringer som diskuteres her, er knyttet til at KI har utallige anvendelsesområder og et enormt potensial til å skape endringer. Utviklingen skjer dels i samhandling med oss, og dels i interaksjon mellom de tekniske systemene. Teknologien er dermed også kjennetegnet av uforutsigbarhet; det er vanskelig eller umulig å forutse hvilke virkninger teknologien vil ha for enkeltindivider, samfunnet og miljøet.

5. Erkjenne usikkerhet
Samtidig som KI byr på store muligheter, står vi overfor betydelig usikkerhet. Uansett hvor gode intensjonene har vært, kan bruken eller konsekvensene vise seg å bli kontraproduktive eller negative, fordi vi ikke har full kunnskap om hvordan teknologien fungerer eller hvordan den vil bli anvendt. I likhet

med andre muliggjørende teknologier er KI dermed kjennetegnet av *uforutsigbarhet*; det er vanskelig eller umulig å forutse hvilke virkninger teknologien vil ha for enkeltindivider, samfunnet og miljøet. Dette er en problemstilling som også berøres i flere av innspillene til NENT. Usikkerheten i KI-forskning er knyttet til de følgende dimensjonene: a) utviklingen og tilpasningen av systemene, inkludert kvaliteten på grunnlagsdata; b) bruken av systemene og deres konsekvenser for enkeltpersoner, dyr, miljøet og samfunnet; c) de verdiene som eksplisitt eller implisitt bygges inn i systemene og hvordan disse påvirker utfall, eller sett i en større sammenheng enkeltpersoner, dyr, miljøet og samfunnet; og d) konsekvensene av ikke å utvikle teknologien.

Uforutsigbarheten i teknologiutviklingen gir opphav til en mye omtalt utfordring som KI deler med andre muliggjørende teknologier, og som omtales som «Collingridges dilemma». Dilemmaet henviser til hvordan utviklingen vanskelig kan styres i en tidlig fase, fordi det fulle omfanget av konsekvensene ofte er uklart før samfunnet har tatt kunnskapen og teknologien i bruk. Da er det gjerne for sent med regulering, ettersom det har vist seg vanskelig å holde tilbake teknologi som er utviklet, eller trekke tilbake teknologi som allerede er tatt i bruk. De spørsmålene som reises i forbindelse med uforutsigbarhet, er blant annet om vi bør rette søkelyset mot langsiktige og fremtidige konsekvenser eller mot mer umiddelbare effekter, hva slags usikkerhet som er aktuell, og hvordan vi kan håndtere denne usikkerheten. Både EU og Norges forskningsråd har lansert «Responsible Research and Innovation», RRI, (Ansvarlig forskning og innovasjon) som et forsøk på å møte slike utfordringer. I litteraturen om RRI har svaret ofte vært å peke på et framoverskuende ansvarsbegrep. Dette handler om å ta ansvar tidlig i et forskningsløp og sikre at gode valg kan tas i den videre prosessen, dels ved å foregripe og vurdere mulige konsekvenser, og dels ved å bygge opp et apparat for å håndtere dem.

Forskningsetiske retningslinjer for naturvitenskap og teknologi legger på sin side vekt på at forskningen også har et ansvar for å formidle usikkerhet ved egen forskning og vurdere risikoene knyttet til implikasjonene av den egne virksomheten:

> Forskeren skal få klart fram usikkerhet i egen forskning og vurdere risiko som følge av forskningsfunn.
> Forskeren skal få fram den graden av sikkerhet og presisjon som kjennetegner forskningsresultatene. Spesielt skal forskeren være nøye med å klargjøre funnenes relative sikkerhets- og gyldighetsområde.
> I tillegg til å framstille kunnskap kritisk og i kontekst, skal forskeren bestrebe seg på å påpeke eventuelle risiko- og usikkerhetsmomenter som kan ha betydning for fortolkning og eventuelle anvendelser av forskningsfunnene. Å formidle et klart bilde av den relative sikkerheten og gyldigheten av kunnskapen er en del av forskerens etiske ansvar og streben etter objektivitet. Der det er mulig, bør forskeren også benytte seg av egnete metoder for å framstille usikkerhet i forskningen. Forskningsinstitusjonene har en plikt til å formidle slike metoder til sine ansatte og studenter (retningslinje 8).

NENT ser et behov for systematiske studier av de risikoenesom er forbundet med utviklingen av KI. Det er viktig at både forskere og politiske beslutningstakere erkjenner usikre, men mulige konsekvenser, og også *ukjente* ukjente, dvs. fremtidige konsekvenser vi ennå ikke kjenner til. Myndigheter og forskningsfinansierende institusjoner bør legge til rette for tverrfaglighet i forskningen, for slik å bedre erkjenne uforutsigbarhet og minimere usikkerhet der det er mulig.

6. Sikre bred involvering

Mange av de mulighetene og risikoene som er forbundet med utviklingen av KI, er usikre og vanskelige å identifisere, men enkelte kan allerede pekes ut og sannsynliggjøres. Flere av de utfordringene som oppstår med KI, er relevante også for andre muliggjørende teknologier, slik som bioteknologi og nanoteknologi. Et felles trekk ved slike teknologier er deres brede potensial til å forandre samfunnet ved hjelp av de mulighetene de gir til å etablere nye koplinger mellom ulike fagfelt og aktiviteter. På samme måte vil en utvikling av KI kunne gi slike systemer utallige anvendelsesområder i samfunnet. På den ene siden kan forskningen bidra til å løse store samfunnsutfordringer innen kjerneområder som helse, energi, klima og sikkerhet. På den annen side kan de imidlertid gi opphav til bekymring om risiko for mulig misbruk og uønskede konsekvenser. De fleste innspillene til NENT understreker

de store mulighetene som er knyttet til KI. Mange viser imidlertid også til mulige negative konsekvenser av utviklingen, men samlet sett fremstår miljøene som optimistiske, og innspillene gjenspeiler i mindre grad de bekymringene som er fremmet internasjonalt av flere forskere de siste årene.

Universitetet i Bergen skriver følgende:

> Den teoretiske forskningen innen KI er av en slik karakter at den kan ha anvendelser innen mange felt. Det gjelder som for matematikken og kjernefysikken at kunnskapen kan ha anvendelser innen f.eks. helse, men også bidra til omdiskuterte anvendelser som våpenutvikling eller til bruk til formål som opplagt medfører skade for enkeltpersoner og samfunn.

NMBU understreker behovet for tverrfaglighet og en bred og involverende debatt:

> Forskningen bør være tverrfaglig fordi KI også kan handle om mennesket og konsekvenser for samfunnet, og det kan være behov for å etablere etiske kjøreregler for utviklingen av ny teknologi.

I forskningsetikken finner vi forsøk på å møte utfordringene knyttet til samfunnsovergripende konsekvenser. *Forskningsetiske retningslinjer for naturvitenskap og teknologi* understreker forskernes selvstendige samfunnsansvar:

> Forskningen har et selvstendig ansvar for egen rolle i samfunns-utviklingen.
> Forskere og forskningsinstitusjoner skal bidra til en felles kollektiv kunnskapsbygging og til å løse store utfordringer som verdenssamfunnet står overfor (retningslinje 1).

Den første retningslinjen innebærer at forskere må reflektere kritisk over og redegjøre for sin egen rolle i teknologi- og samfunnsutviklingen. NENT mener det er viktig at fagmiljøene også selv kritisk vurderer visjonene bak KI-forskningen og hva som er legitime og mindre legitime formål. Regjeringens strategi for KI vil trolig langt på vei definere hva visjonen med

KI vil bli nasjonalt, og dermed legge føringer for forskningens retning. I mange tilfeller vil formålet i stor grad være forutbestemt av en oppdragsgiver, og den videre bruken vil ofte være bestemt av andre aktører. I dette rommet kan det likevel ligge et betydelig ansvar hos forskeren, i den grad det foreligger en mulighet til å påvirke hvorfor og hvordan KI-systemer utvikles.

Forskere har også et ansvar for å kommunisere den risikoen som følger av forskningsfunn. Føre var-prinsippet kan være relevant i håndteringen av risiko under vitenskapelig usikkerhet. Dette er formulert på følgende måte i *Forskningsetiske retningslinjer for naturvitenskap og teknologi*:

> Forskeren skal bestrebe seg på å bidra til å følge føre var-prinsippet
> Der det foreligger plausibel, men usikker kunnskap om at en teknologisk anvendelse eller en utvikling av et forskningsfelt kan føre til etisk uakseptable konsekvenser for helse, samfunn eller miljø, skal forskerne innenfor det aktuelle forskningsfeltet bestrebe seg på å bidra med kunnskap som er relevant for å følge føre var-prinsippet. Dette innebærer at forskeren skal samarbeide med andre relevante parter i å følge føre var-prinsippet. Føre var-prinsippet defineres her på følgende måte: «Når menneskelige aktiviteter kan føre til moralsk uakseptabel skade som er vitenskapelig rimelig, men usikker, skal man foreta handlinger for å unngå eller minske slik skade.» Dette prinsippet er viktig for store deler av den naturvitenskapelig-teknologiske forskningen, og forskere har et medansvar for å legge forholdene til rette for vurderinger knyttet til føre var-prinsippet og bidra til å unngå eller minske skade (retningslinje 9).

Føre var-prinsippet gjelder ikke i tilfeller med full usikkerhet, bare der det foreligger «plausibel, men usikker kunnskap». På KI-feltet er det stor usikkerhet, og det er også uenighet om bestemte negative konsekvenser faktisk vil materialisere seg i fremtiden, særlig når det gjelder fullt autonome systemer. Utviklingen av spesifikk KI har imidlertid mange allerede kjente konsekvenser, og risikoene er i mange tilfeller tilstrekkelig sannsynliggjort til å aktualisere føre var-prinsippet. Føre var-prinsippet innebærer at KI-forskere må beskrive og kommunisere den risikoen som er forbundet med utviklingen og bruken av KI på deres forskningsfelt. Hvilke «etisk uakseptable konsekvenser for helse, samfunn og miljø» som vektlegges, vil imidlertid

variere i tråd med hvilket etisk perspektiv og hvilke verdier og interesser som legges til grunn. Press fra ulike aktører kan påvirke algoritmene, uten at dette er gjort til gjenstand for en faglig eller politisk vurdering. For eksempel kan overvåkingssystemer basert på KI oppfattes både som et mulig gode og som en risiko. I et forsvars- og sikkerhetsperspektiv kan overvåking betraktes som et gode som forhindrer kriminalitet og advarer samfunnet mot mulige farer, men ut fra et ståsted som vektlegger personvern, kan slik overvåking også anses som en trussel mot enkeltpersoners integritet.

De som blir mest påvirket av de beslutningene som tas, må også sikres en stemme i beslutningsprosessene. Myndigheter og forskningsinstitusjoner bør derfor legge til rette for at innbyggerne kan bli bredt involvert i en debatt om hva som skal være forskningens formål, innretningen på forskningssatsninger og bruken av forskningen.

C) Stordata
Sammen med datakraft og algoritmer har utviklingen av spesifikk KI i stor grad vært drevet frem av stordata, som også kan inkludere personopplysninger. Det tredje setter med utfordringer NENT mener det er viktig å rette oppmerksomheten mot, er knyttet til stordata innenfor KI-forskningen. Stordata gir opphav til nye utfordringer knyttet til personvern og beskyttelse av enkeltpersoner i forbindelse med forskning. Stordata reiser også andre forskningsetiske spørsmål som vi skal ta opp i det følgende, blant annet knyttet til skjevheter i datamaterialet, datakvalitet og eierskap og tilgang til data.

7. Sikre personvern og hensynet til enkeltpersoner
Beskyttelse av data som inneholder personlige opplysninger, kan by på egne utfordringer i utviklingen og bruken av KI. Selv om det benyttes anonymiserte data i analysene vil sammenstillinger med andre data likevel kunne avdekke sensitive opplysninger eller avsløre enkeltpersoner, og dermed utgjøre personlige data. Innhenting og bruk av data som inkluderer persondata kan utfordre kravet om informert samtykke. Ved innsamling og nye sammenstillinger av store datamengder er det en særlig risiko for at personlig informasjon kan brukes på måter vi ikke kjenner til (fordi formålet også er ukjent for forskeren på tidspunktet for innsamlingen), og som vi

kanskje ikke ønsker.

Hensynet til individer og grupper som på ulike måter er involvert i forskningen eller påvirkes direkte, er regulert blant annet i personvernreglene (dvs. EU-regelverket og den supplerende norske personopplysningsloven). Personvernreglene er en viktig rettesnor for forskere, men kan ikke alene gi svar på de mange utfordringene som forskere vil stå overfor i forbindelse med håndteringen av personopplysninger. I Norge utgjør *Forskningsetiske retningslinjer for samfunnsvitenskap, humaniora, juss og teologi* (NESH 2016) det sentrale verktøyet for å utdype det etiske ansvaret overfor forskningspersoner og andre som berøres av forskningen. Implementeringen av personvernreglene i norsk rett innførte en rekke grunnleggende prinsipper som må overholdes for at behandling av personopplysninger skal være lovlig. Ett av disse er prinsippet om *dataminimering*. Av dette følger det at man ikke skal benytte flere personopplysninger enn det som er nødvendig for å oppfylle formålet med den behandlingen som søkes gjennomført, og opplysningene må i tillegg være adekvate og relevante for behandlingen. For de som arbeider med KI, kan det imidlertid by på utfordringer å begrense den mengden opplysninger som behandles, fordi utvikling og bruk av KI vanligvis har behov for store mengder data for å lære opp systemene. For å vurdere hva som er nødvendig, adekvat og relevant, må forskeren ha klart for seg hva som er formålet den behandlingen som ønskes gjennomført.

Et annet grunnleggende personvernprinsipp er nettopp at databehandlingen skal være *formålsbegrenset*. Behandling av personopplysninger kan ikke skje uten at det foreligger et berettiget, spesifikt og uttrykkelig angitt formål. At formålet skal være klart og spesifikt innebærer at det må være konkret beskrevet. Fordi dette formålet er førende for oppfyllelsen av en rekke av de andre personvernprinsippene, er vage og omfattende angivelser av formålet ikke tillatt. For eksempel skal personopplysninger slettes når formålet med behandlingen er oppfylt. Dersom formålet er angitt å være utvikling av KI, vil det imidlertid være vanskelig å imøtekomme dette kravet. Kravet til formålsbegrensning vil imidlertid trolig kunne oppfylles ved at man angir hva slags KI som skal utvikles, og hvilke oppgaver man antar at dette systemet vil kunne utføre. Det vil uansett kunne bli utfordrende

å avgjøre hvorvidt kravene til dataminimering og formålsbegrensning er oppfylt i slike sammenhenger. I sin rapport om KI og personvern bemerker Datatilsynet at man ved utvikling av KI bør søke å begrense treningsdataene ved oppstart, for deretter å utvide datasettet når man i større grad vet hva man har behov for. Artikkel 29-gruppen, en rådgivende gruppe innenfor EU, peker i sin uttalelse om automatiserte avgjørelser på viktigheten av at den behandlingsansvarlige innfører rutiner og systemer som sørger for at de benyttede personopplysningene til enhver tid er korrekte og oppdaterte. Hensikten med disse prinsippene som ligger bak reguleringen, er at behandling av personopplysninger kun skal skje når det er nødvendig, for slik å begrense inngrepet i den enkelte registrertes personvern. Ved bruk av store datamengder for utvikling av KI er det viktig å ha dette for øyet, og vurdere om datamengden kan begrenses uten at dette går på bekostning av det angitte formålet. Dersom datamengden ikke kan begrenses, må dette valget kunne begrunnes og forklares for å kunne vise at dataminimeringsprinsippet er oppfylt på samme måte som kravene til nødvendighet, adekvans og relevans. Hvis dataene skal gjenbrukes, må utviklerne forsikre seg om at denne bruken er i overensstemmelse med det opprinnelige formålet.

Hovedregelen i forskningsetikken er at personopplysninger ikke skal samles inn, behandles eller deles uten informert samtykke. Når data som er innsamlet til andre formål, gjenbrukes på nye og uventede områder, bør samtykket oppdateres der dette er mulig. Utfordringer kan også oppstå når data som i utgangspunktet er anonymiserte, blir sammenstilt på nye måter. I sine forskningsetiske vurderinger av informasjon og samtykke har forskere har et ansvar for å vurdere opplysningenes offentlighet, informasjonens sensitivitet, de berørtes sårbarhet og forskningens interaksjon og konsekvenser (NESH 2019).

8. Kvalitetssikring

I forbindelse med KI-forskning kan det være særlig grunn til å stille kritiske spørsmål ved dataenes kvalitet, sannferdighet og relevans, fordi vi ikke alltid kjenner kildene til dataene og fordi metadata kan mangle eller være usikre. Skjevheter i materialet, egenskaper ved analyseverktøyet og menneskelig fortolkning øker mulighetene for logiske feilslutninger og diskriminerende beslutninger. Dette gir grunnlag for usikkerhet i fortolkninger og

beslutninger som er basert på KI. I de senere årene har vi sett flere eksempler på hvordan data kan gi opphav til urimelige avgjørelser. Da Amazon i 2018 forsøkte å etablere en objektiv ansettelsesprosess ved hjelp av KI, viste denne seg å gi utslag i kjønnsdiskriminerende beslutninger, fordi datasettene favoriserte menn.

> KI må trene seg opp på virkelige data. Teknikker som dyp læring fungerer best med mye data. Da er vi prisgitt datakvaliteten, som ikke alltid er god (Norsk Regnesentral).

NENT mener at for å sikre etterprøvbarhet og kvalitet er det vesentlig at forskere og forskningsinstitusjoner legger til rette for at datakilder skal være åpne og allment tilgjengelige. Samtidig bør usikkerhetsfaktorene og begrensningene ved forskningen erkjennes og kommuniseres.

9. Rettferdig tilgang til data
Utviklingen av KI-teknologi kan medføre at et fåtall individer, selskaper eller forskningsgrupper får mulighet til å dominere dette feltet.

> Den største bekymringen er imidlertid at KI ser ut til å bli dominert av noen enkeltaktører som Facebook, Google, Amazon og noen andre. Den virkelig godt fungerende KI er avhengig av store mengder data og datakraft. Firmaer som f.eks. Facebook har enorme mengder data og datakraft andre aktører umulig kan matche (CAIR, UiA).

NENT ser en risiko for at store deler av forskningsinnsatsen innen KI unndrar seg de kravene til åpenhet som gjelder for forskning ellers, slik de blant annet er nedfelt i FAIR-prinsippene, for eksempel med henvisning til behovet for hemmelighold til beskyttelse av konkurransefortrinn. Forskningsetisk sett er det vesentlig å sørge for at forskningen, inkludert data og resultater, som hovedregel gjøres tilgjengelig for alle. Som formulert i *Forskningsetiske retningslinjer for naturvitenskap og teknologi* innebærer forskningsetikkens krav til åpenhet at forskningsresultater, metoder og data deles og offentliggjøres, både for å legge til rette for kvalitetssikring, opprettholde tilliten til forskningen og sikre at resultatene kommer samfunnet til gode (jf. retninglinje 3, 4 og 17).

Manglende deling av data er problematisk av flere grunner. For det første, dersom det bare er forskere innen et fåtall privilegerte bedrifter som får adgang til å analysere store datasett, vil det bli umulig for utenforstående å reprodusere og evaluere resultatene deres. For det andre kan forskere som er tett koblet på private selskaper, ha motivasjoner og interesser som vil påvirke både hvilken forskning som prioriteres og resultatene av den. Bedrifter som utfører evalueringer eller forskning, gjør dette med et kommersielt mål for øyet. Forskere som har hensiktsmessig kompetanse og tilgang til de riktige dataene, kan bidra til å frembringe et bedre kunnskapsgrunnlag som igjen kan komme samfunnet til gode på en bredere måte.

NENT mener at myndigheter og forskningsinstitusjoner bør legge til rette for allmenn tilgang til data. De bør sørge for åpenhet om hvem som skal ha eierskap til teknologi, infrastruktur og data, hvilke forskningsområder som blir prioritert og hvorfor, og hvem som kan forventes å ha nytte av forskningsinnsatsen.

Avslutning

De ni punktene i denne betenkningen er ment å tjene som et utgangspunkt for refleksjon, veiledning og diskusjon i forskningsmiljøene. De er også utarbeidet for aktører som finansierer og legger til rette for KI-forskning, eller som tar i bruk KI. Fordi utviklingen av dette forskningsfeltet er preget av høy hastighet og usikkerhet, bør denne betenkningen etter en tid bli gjenstand for nye vurderinger og revisjoner. NENT ønsker å fortsette dialogen med fagmiljøene om de forskningsetiske utfordringene ved KI-forskning, og dermed legge til rette for etisk god og ansvarlig KI-forskning i Norge.

De nasjonale forskningsetiske komiteene
Kongens gate 14, 0153 Oslo
Tlf.: 23 31 83 00
www.etikkom.no
ISBN: 978-82-7682-097-3