Økende utbredelse av kunstig intelligens kan medføre etiske dilemmaer. Illustrasjon: hafakot/Shutterstock.

Etiske problemstillinger ved kunstig intelligens:

Respekt for menneskelig autonomi

Ieva Martinkenaite og Geir Egil Dahle Øien

Den stadig økende utbredelsen av kunstig intelligens-teknologi (KI) har medført en økende diskusjon om potensielle risiki og etiske dilemmaer knyttet til utvikling og bruk. De største bekymringene som diskuteres globalt er forbundet med umoralsk og uansvarlig bruk av KI, som potensielt kan skade menneskeheten, true demokratiet, forsterke sosioøkonomiske ulikheter, og sette grunnleggende menneskerettigheter og verdier i fare.

Ieva Martinkenaite (f. 1980) er blant nøkkelpersoner i Telenor Group som bidrar til å bygge AI-forskning og innovasjonsøkosystemer i Norge. Hun er visepresident i Telenor Research som leder et team av dataforskere og ML-ingeniører og er styremedlem i Norwegian Open AI-Lab. I løpet av 2018–2020 representerte Ieva Telenor i EU-kommisjonens ekspertgruppe om KI, som ga etiske retningslinjer for pålitelig AI. Ieva har doktorgrad i strategi (2015) fra Handelshøyskolen BI.

Geir Egil Dahle Øien (f. 1965) er siv.ing. (´89) og dr.ing. (´93) fra NTH. Han ble professor i informasjonsteori ved NTNU i 2001, og var dekan ved NTNU 2009–2019. Øien leder p.t. prosjektet «Fremtidens teknologistudier», samt Porteføljestyret for muliggjørende teknologier i Norges forskningsråd. Han har deltatt i en rekke prosjekter finansiert av Norges Forskningsråd og EU, veiledet over 20 ph.d.-kandidater, og vært medforfatter på rundt 150 vitenskapelige artikler. Han er medlem av NTVA og DKNVS.

Vi lever i en tid preget av enestående endring. Covid-19-pandemien har rystet både økonomien, politikken og forutsetningene for den langsiktige samfunnsutviklingen. Utbruddet har også vist at mange av samfunnsendringene er drevet av teknologiske fremskritt, blant annet innen stordata og kunstig intelligens (KI).[1] Banebrytende teknologisk innovasjon både flytter grensene for hva vi kan utrette, og endrer metodene vi bruker for å løse oppgaver. Kjente eksempler fra dagliglivet er mobiltelefoner med innebygde databehandlingsbrikker basert på nevrale nettverk – som muliggjør ansiktsgjenkjenning, prediktiv skriving, naturlige språkgrensesnitt for «tastaturet», produktanbefalinger og personlig nyhetsfeed på Facebook, «raskeste vei»-anbefalinger på Google Maps og Teslas autopilot.

Utbredelsen av KI-teknologier har også medført økende diskusjon om potensielle risiki knyttet til utvikling og bruk av kunstig intelligens. De største bekymringene er forbundet med umoralsk og uansvarlig bruk av KI, som potensielt kan skade menneskeheten, true demokratiet, forsterke sosioøkonomiske ulikheter og sette grunnleggende menneskerettigheter og verdier i fare. Massiv overvåking av innbyggere, vurdering av mennesker gjennom poengsummer i et «sosialt kredittsystem», skjulte KI-systemer som manipulerer menneskers oppfatninger i sosiale medier, og utvikling og bruk av KI for autonome våpensystemer er eksempler på fenomener som har utløst stor medieoppmerksomhet, og også kritikk fra ledere og andre verden over. En rekke regjeringer og institusjoner over hele verden har publisert KI-etiske prinsipper som styrer utvikling og bruk.[2]

Samtidig kan det være sterke etiske argumenter *for* å utvikle og ta i bruk KI-teknologi. For eksempel kan mer effektiv rekognosering og målidentifisering i krigføring i beste fall medføre mindre ødeleggelse og færre dødsfall. Dette indikerer at det kan være gode argumenter for å forske på og utvikle slik teknologi, selv der det opplagt må settes grenser for bruk. Jus er et annet eksempel: Det er vist at maskinassisterte juridiske dokumentgjennomganger er raskere, mer nøyaktige og billigere enn hva mennesker kan utføre.[3] Kan det ikke da etisk sett være å foretrekke å bruke slik teknologi i juridisk praksis?

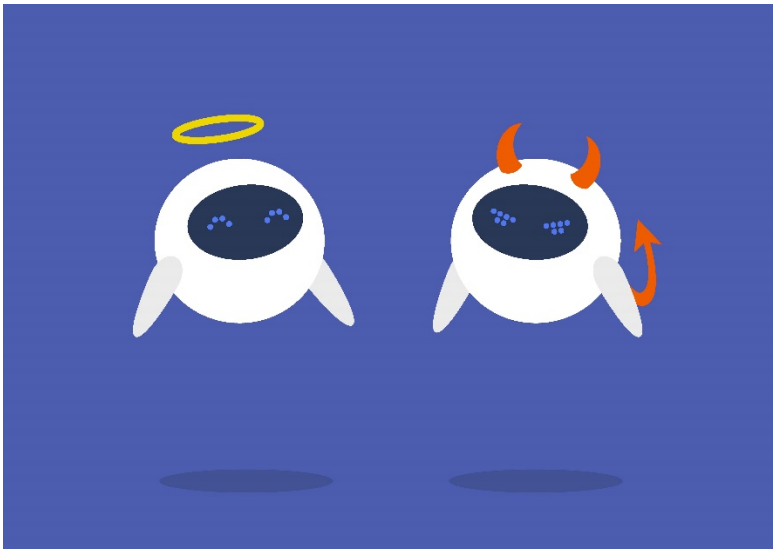
Det ovenstående indikerer at debatten rundt KI og etikk er global, stadig aktuell og mangefasettert. Med denne artikkelen ønsker vi å peke på noen av de viktigste etiske utfordringene, gi eksempler (også i en norsk kontekst), og informere leseren om forskjellige perspektiver i debatten, med mål om å fremme inkludering og bred dialog.

---

[1] I bredere forstand kan kunstig intelligens defineres som vitenskap med mål om å gjøre maskiner smarte. KI er en samlebetegnelse for datasystemer som er i stand til å sanse miljøet sitt, forstå, handle og lære av erfaring for å løse problemer, f.eks. å gjenkjenne objekter i bilder, oppdage feil på oljeinstallasjoner offshore eller automatisk oversette tale til tekst (Purdy & Daugherty, 2017). Maskinlæring (ML) er et underfelt av AI som lar datamaskiner lære direkte fra eksempler, data og erfaring uten å være eksplisitt programmert. ML lar datamaskiner lære mønstre, finne sammenhenger og identifisere avvik fra store datamengder (også kalt «Big Data»).

[2] Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. & Srikumar, M. *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*. Berkman Klein Center for Internet & Society, 2020.

[3] Grossman, M.R. &. Cormack, G.V. (2011). *Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review*. Richmond Journal of Law and Technology, 17(3) og Daniel Martin Katz, Michael J. Bommarito II and Josh Blackman (2017). *A General Approach for Predicting the Behavior of the Supreme Court of the United States*, PLoS ONE, *12*. 2017. (https://doi.org/10.1371/journal. pone.0174698).



Figur 23.1 Roboter kledt som engler og djevler. Illustrasjon Nadia Snopek/Shutterstock.

KI-etikk: Hva er det, og hvorfor er det et så *hot* tema?

Vi forbinder gjerne etikk og etiske vurderinger med spørsmål som «Hva er en god handling?», «Hva er verdien av et menneskeliv?», og «Hva er rettferdighet?». *KI-etikk* fokuserer på normative spørsmål – dvs. hva vi bør (eller har lov til å) gjøre – i utviklingen og bruk av KI-teknologier, med mål om å forbedre individuell og kollektiv livskvalitet og velferd. Ifølge EU-kommisjonens *Etiske retningslinjer for pålitelig KI*[4] må etisk KI sikre overholdelse av følgende fire prinsipper:

- Respekt for menneskelig autonomi,
- Forebygging av skade,
- Rettferdighet, og
- Forklarbarhet.

KI-etikk omhandler med andre ord grunnleggende moralske og fysiske rettigheter for enkeltpersoner – rettigheter som oppstår i kraft av personenes menneskelighet, uavhengig av deres juridiske status. Som terminologien antyder, kan vi altså assosiere etisk bruk av KI med å bruke KI for å fremme menneskelig verdighet og ha en positiv innvirkning på samfunnet. Positive eksempler varierer fra dyplæringssystemer[5] som er i stand til å forutsi strukturer av SARS-CoV-2-proteiner[6], via KI-baserte medisinske verktøy for diagnostisering av lungekreft i tidlig stadium,[7] stemmeassistenter for

[4] Ethics Guidelines for Trustworthy AI (2019). Independent High-Level Expert Group on Artificial Intelligence set by the European Commission. (https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai)

[5] Dyplæring er et underfelt av maskinlæring der kunstige nevrale nettverk – algoritmer inspirert av den menneskelige hjerne – lærer av store mengder data. På samme måte som vi lærer av erfaring, vil en dyplæringsalgoritme utføre en oppgave gjentatte ganger, og hver gang tilpasse løsningen litt for å forbedre resultatet. Begrepet «dyplæring» viser til at nevrale nettverk har flere ulike (dype) lag som muliggjør læring. For mer om dyplæring, se følgende definisjon og eksempler.

[6] https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

[7] https://www.nature.com/articles/d41586-020-03157-9

synshemmede[8] og bruk av roboter i smart industriproduksjon, til algoritmer som kan hjelpe til med å forutsi orkaner.[9]

Men KI-teknologier kan også bli *underbenyttet* (eng. *underused*), *forsettlig misbrukt* (eng. *misused*) eller *utilsiktet brukt* (eng. *overused).* Alle tre scenarioene betraktes som uetiske (noen ganger til og med ulovlige), da de implisitt kan utgjøre en risiko for samfunnet. For eksempel kan underinvestering i automatisering av manuelle og kjedelige arbeidsprosesser føre til urealiserte fordeler av ny teknologi for samfunnet. Alt fra epost-svindel og falske nyheter til fullskala cyber-krigføring kan forsterkes gjennom ondsinnet bruk av KI. Og skjult manipulasjon av menneskelig adferd i f.eks. shopping, promotering av hatefulle ytringer på sosiale medier, og normative poengsystemer for å klassifisere borgere er alle eksempler på risiki forbundet med forsettlig misbruk eller overbruk av KI-teknologier.

EUs lovgivere har valgt en tilnærming til KI-etikk som er forankret i grunnleggende rettigheter, som nedfelt i EUs verdier og internasjonal menneskerettighetslov, med et mål om å sette en global standard. Dette er kanskje plausibelt, men man bør samtidig huske at det som betraktes som «moralsk» eller «etisk» (og dermed sosialt akseptabel) bruk av denne nye teknologien kan være forskjellig både på tvers av kulturer, grupper eller individer, og på forskjellige tidspunkter. Dette tilsier at begrepet KI-etikk har en mangefasettert, sosial og situasjonsmessig karakter.

Selv om det er forskjellige holdninger til denne nye teknologien, har etiske spørsmål om KI skapt medieoverskrifter internasjonalt, kommet høyt opp på den (geo-)politiske agenda hos verdens ledere, og utløst offentlig debatt over hele verden. Hvorfor? Det er minst fire grunner til det:

1. **Økende grad av delegering av beslutningsmakt –** til (delvis) autonome KI-systemer (basert på maskinlæring) som blir stadig mer effektive, og som har potensial til å både drastisk endre folks hverdag, og bidra til viktige beslutninger innen bl.a. mobilitet, handel, helse, utdanning og rettshåndhevelse. For eksempel bruker vi Google Maps for å navigere i travle gater, og vi kjøper ting basert på KI-genererte anbefalinger på kommersielle nettsteder, bruker KI-drevne verktøy i medisinsk diagnostikk, og tilpasser utdanningsplanene våre basert på råd fra KI-algoritmer.

2. **Mangel på gode overvåkingssystemer** for KI-/ML-teknologi og -anvendelser – eksemplifisert ved bruk av «umoden» KI med lite tilsyn og ansvarlighet i ulike høyrisiko-situasjoner. Vi kan nevne prediktive politialgoritmer med utilsiktet slagside (bias) i USAs strafferett[10], svindelannonser basert på «deep fakes»[11], og ansiktsgjenkjenning brukt i algoritmer for massiv overvåking i Kina.[12]

3. **Begrenset generell kunnskap om KI** i samfunnet, og mangel på demokratisering (tilgjengelighet og rimelighet) av KI-teknologi – noe som forverrer – noen ganger også

[8] https://medium.com/@rmerrett/the-potential-for-voice-interfaces-in-assisting-the-blind-6506d3da5e87
[9] https://doi.org/10.1029/2020GL089102
[10] https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/
[11] https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them
[12] Feldstein, Steven. (2019) *The Global Expansion of AI Surveillance. Carnegie Endowment for International Peace*. https://carnegieendowment.org/files/WP-Feldstein-AISurveillance_final1.pdf

ubegrunnet – bekymringer over KI som en trussel, både i mediene og i den politiske og offentlige debatten.

4. **Ujevn fordeling av makt og utviklingskapasitet knyttet til KI** – med kompetanse, ressurser og kapasitet konsentrert i USA og Kina (og til en viss grad i Storbritannia, Canada, Israel, Singapore), fører til geopolitiske bekymringer.

Figur 23.2 Dyp læring og maskinlæring. Illustrasjon: kentoh/Shutterstock.

Etiske problemstillinger knyttet til KI, og internasjonal respons

Vi kan definere sju overordnede temaer i den globale debatten om «etisk KI». Disse er:

1. **Menneskelig autonomi:** *Er det mennesker i loopen?* Med økende delegering av makt til datadrevne, (delvis) autonome systemer om kritiske beslutninger i folks liv, dukker det opp spørsmål om menneskelig verdighet og frihet, menneskelig tilsyn med beslutningene og selvbestemmelse. Blant de største bekymringene er: *deep-fake* videoer med sikte på å manipulere og styre folks meninger (f.eks. ved valg), ansiktsgjenkjenning for masseovervåking, og behovet for menneskelig verifisering av autonomt klassifiserte medisinske data, som grunnlag for beslutninger rundt behandling av f.eks. kreftpasienter. Én forventning er at mennesker som samhandler med KI-systemer skal kunne beholde full og effektiv selvbestemmelse, og fremdeles være i stand til å delta i demokratiske prosesser. KI-systemer bør altså ikke uberettiget underordne, tvinge, lure, manipulere eller påvirke mennesker. I stedet bør de utformes for å øke, utfylle og styrke kognitive, sosiale og kulturelle menneskelige ferdigheter.

2. **Rettferdighet og ikke-diskriminering:** *Er det rettferdig***?** Maskinlæringsmodeller trenes typisk opp ved hjelp av store datamengder. Disse dataene kan inneholde skjevheter og fordommer, og til slutt gjøre KI-avgjørelser urettferdige og diskriminerende. Dette er kjent som «skjevhet» eller «bias», og kan oppstå både i utvalget av treningsdata, i selve datakilden eller på grunn av feil i algoritmer. Eksempler på bekymringer man har er at algoritmiske skjevheter skal gi feilaktige juridiske beslutninger, og at helsebeslutninger skal bli påvirket av søppeldata og dårlig utformede KI-systemer. Eksperter peker også på at det kan ligge implisitte forstyrrelser i det faktum at KI-utviklere i dag domineres av velutdannede og økonomisk uavhengige hvite menn. Ulik fordeling av KI-kunnskap og -investeringer kan forsterke geopolitisk og makroøkonomisk ulikhet, og dermed være diskriminerende på lang sikt. Noe av denne kraftige

teknologien vil være så avansert (og eventuelt dyr å utvikle) at den kun kan brukes av dem med makt, midler og tilgang, mens andre står i fare for å forbli totalt utenfor. Et hovedspørsmål blir da: Kan forskjellene mellom rike og fattige, mektige og ikke-mektige, øke ytterligere? Og kan folk i land med dårlige rettsstater og mye fattigdom på forskjellig vis bli ofre for dem som har råd og mulighet til å bruke kraftig KI-teknologi, både i eget land og annetsteds?[13]

For å redusere disse svakhetene krever politiske beslutningstakere i flere land nå større grad av mangfold og inkludering i KI-arbeidsstyrken, og investeringer i mer variert rekruttering til feltet. Forsknings- og forretningsorganisasjoner bes også om å utforme regler for hvordan man kan forhindre, oppdage og eliminere diskriminerende data og algoritmiske skjevheter som en del av arbeidet med KI-etiske prinsipper. *Personvern-rettferdighets-paradokset*[14] peker samtidig på vanskeligheter med å teste KI-systemer mot rettferdighetsparametere når personopplysninger som kjønn, rase, religion ikke er tillatt å bruke i henhold til GDPR. Begrepet *fullstendighet* er også nært knyttet til rettferdighetsbegrepet: KI sies å ikke være «fullstendig nok» dersom den savner tilgang til visse data som hindrer teknologien i å være effektiv i sin oppgave. Ett eksempel er Amazons rekrutteringsverktøy, som hovedsakelig ble trent på mannlige ansatte, og endte opp med å diskriminere mot kvinnelige kandidater.

3. **Åpenhet og tolkbarhet (eng. interpretability):** *Hvordan forstå hva som skjer inne i KI-teknologien?* Rask utvikling av ML forsterker «black box»-problematikken, der det er uklart *hvordan* og *hvorfor* en algoritme når frem til sin prognose eller anbefaling – eller i tilfellet roboter, hvorfor de handler slik de gjør. Økende delegering av beslutningsansvar og kontroll overlatt til KI-systemer som er vanskelige å spore, forklare og kritisk evaluere utgjør en betydelig etisk bekymring, også fordi det kan undergrave befolkningens tillit til ny teknologi. Mange er samtidig enige om at full teknisk gjennomsiktighet i ML-systemer er vanskelig, og kanskje heller ikke alltid rimelig å forlange. Dette gir behov for mer forskning på *forklarbar* KI.[15] Det anbefales blant annet å dokumentere hvordan data er blitt samlet inn og merket, hvordan arbeidsflyt i maskinlæringssystemer er blitt implementert, og hvordan algoritmer kommer med sine anbefalinger. Politiske beslutningstakere understreker behovet for å sikre at folk er informert slik at de kan utfordre resultatene av KI-systemer, spesielt i situasjoner der grunnleggende rettigheter kan settes i fare.

4. **Ansvar:** *Hvem er ansvarlig når noe går galt?* Utvikling og bruk av KI-systemer bør ikke frigjøre mennesker fra ansvar når viktige beslutninger eller handlinger blir delegert til maskiner. Ansvarlighet innebærer at juridiske personer til enhver tid bør beholde kontrollen over, og

[13] Forfatterne vil takke Henrik Syse for denne kommentaren.

[14] https://business.blogthinkbig.com/is-your-ai-system-discriminating-without-knowing-it-the-paradox-between-fairness-and-privacy/

[15] De fleste av de siste AI-gjennombruddene kan tilskrives dyplæring (Deep Learning), men til tross for deres imponerende ytelse har DL-modeller ulemper, hvor noen av de viktigste er a) mangel på gjennomsiktighet og tolkning, b) manglende robusthet og c) manglende evne til å generalisere til situasjoner utover tidligere erfaringer. *Explainable AI* (forklarbar KI) tar sikte på å avhjelpe disse problemene ved å utvikle metoder for å forstå hvordan «black box»-modeller gir spådommer og hva som er deres begrensninger. Oppfordringen til slike løsninger kommer fra forskningsmiljøet, industrien og politiske beslutningstakere på høyt nivå, som er bekymret for potensialet ved å distribuere AI-systemer til den virkelige verden når det gjelder effektivitet, sikkerhet og respekt for menneskerettighetene.

ansvaret for, maskiners oppførsel. Regjeringer oppfordres til å avklare hvem som har ansvaret for skader forårsaket av uønsket oppførsel fra autonome systemer – KI-utviklerne, distributører av teknologien, eller brukerne av KI. Det er behov for effektive risikovurderings- og skadebegrensningssystemer.

5. **Sikkerhet:** *Er det trygt?* Sikkerhet, pålitelighet og intern robusthet for KI-systemer må testes og kvalitetssikres før KI-produkter lanseres i markedet. Målet er bl.a. å sikre at produktene ikke truer menneskelig, kroppslig og mental integritet, og at KI-systemene er robuste mot fysiske angrep og nettangrep. Psykologisk manipulering av mennesker gjennom «skjult» bruk av KI-teknologi (for eksempel forsterkningslæring for å avdekke det mest optimale tidspunktet for å selge ulike produkter) gir grunn til bekymring – spesielt siden slike verktøy kan brukes til å skape avhengighet av sosiale medier, øke materielt forbruk, og til slutt påvirke brukernes mentale helse og sosiale liv. De «røde linjene» forbundet med utvikling og bruk av autonome og dødelige våpen (også kjent som «killing robots»), f.eks. i væpnede konflikter eller målrettet mot personer i sårbare stillinger, blir også sett på med stor bekymring av mange. Dette er én grunn til at politiske beslutningstakere ønsker å innføre beskyttende ordninger som kan gjøre det mulig for mennesker å slå av «funksjoner» for å forhindre at KI-systemer fortsetter handlinger som kan være skadelige. Begrepet *nøyaktighet* er også innført i debatten, i den forstand at bare de mest nøyaktige KI-systemene vil være i stand til å oppnå tillit til at de faktiske resultatene de leverer er riktige. F.eks. når KI skanner et notat fra en lege, må den både kunne lese teksten riktig, forstå den, og ta en riktig avgjørelse basert på hva legen har skrevet, for å kunne oppnå tillit.

6. **Personvern og datastyring:** Mangel på åpenhet om hvordan personopplysninger samles inn, lagres og brukes, og feil bruk av private data, gir også samfunnsmessige bekymringer. KI-teknologi kan muliggjøre målrettet overvåking av borgere, noe som kan være skadelig for personvern og demokrati. Innbyggernes rett til privatliv bør beskyttes, særlig i lys av de omfattende anvendelsene vi nå ser av ansiktsgjenkjenningsteknologi og andre biometriske identifikasjonsmetoder i forbindelse med massiv overvåking uten informert samtykke fra brukerne. EU vurderer å forby ekstern biometrisk identifikasjon i kommende KI-reguleringer.

7. **Sosioøkonomiske implikasjoner**: Mange er bekymret for hvordan utviklingen innen KI vil påvirke jobbene deres. Ett mulig og ikke usannsynlig scenario er da også at KI kan bidra til en mer ulik fordeling av fordeler i samfunnet. Allerede i dag er KI-teknologier modne nok til å erstatte mange rutineoppgaver, og videre utvikling kan medføre at visse yrker blir overflødige eller forsvinner i fremtiden (f.eks. sjåfører, regnskapsførere). For å håndtere endringene teknologien skaper, krever politiske beslutningstakere nå en betydelig investering i livslang læring, utdanning og omskolering av arbeidsstyrken, med særlig fokus på anskaffelse av grunnleggende digitale ferdigheter og kompetanser innen MNT-fag, men også ferdigheter som er komplementære til og ikke kan erstattes av maskiner, som for eksempel kritisk tenkning, empati, kreativitet og entreprenøriell tenkning.

   En kombinasjon av nettverkseffekter, tilgang til enorme mengder data og bruk av KI har dessuten potensial til å styrke dominansen til noen få internettgiganter (som Google, Facebook, Amazon, Alibaba) ytterligere. Dette kan føre til ytterligere monopolisering av data og mer kontroll av menneskers digitale opplevelser. Den nylige konflikten mellom Facebook

og den australske regjeringen[16] viser hvor mektige internettgigantene er blitt. At teknologisk fremgang og KI-distribusjonskapasitet konsentreres til så få aktører gir grunn til bekymring, da det kan skape økt polarisering i samfunnet og uthule demokratiske prinsipper. Fra et bærekraftsperspektiv kan man også peke på høye kostnader og høyt energiforbruk forbundet med å trene de mest sofistikerte dyplæringsmodellene. Dette kan potensielt gjøre det vanskeligere å nå viktige klimaambisjoner hvis KI skal skaleres ytterligere opp.

Hvordan svarer så verdens land på disse utfordringene? Både FN, EU og ulike enkeltland i verden har tatt opp KI-etiske bekymringer i tråd med de ovennevnte punktene, og foreslått ulike måter å håndtere dem på:

- FN har gitt ut AI for Good-serien-serien som den ledende plattformen for internasjonal dialog om KI. Det legges vekt på effektive KI-løsninger som bidrar til å nå bærekraftige utviklingsmål. I 2019 la UNESCO ut på en to-års reise med mål om å utvikle de første globale standardanbefalingene om etikken til kunstig intelligens, med forventninger om et vedtak på UNESCOs generalkonferanse mot slutten av 2021. UNESCOs mål er å utvikle et universelt rammeverk av verdier, prinsipper og handlinger for å veilede sine over 200 medlemsland i utformingen av lovgivning, politikk eller andre virkemidler knyttet til KI.[17]

- EU har også en ambisjon om å sette en global standard for etisk bruk av KI. EU-kommisjonen gjorde etiske og sosiale utfordringer til en integrert del av den europeiske KI-strategien, og har nedsatt en ekspertgruppe som har utarbeidet etiske retningslinjer for pålitelig bruk av kunstig intelligens.[18] Den britiske regjeringen tok ledelsen i å opprette verdens første Center for Data Ethics and Innovation, som vil gi råd til regjeringen om etisk og innovativ bruk av KI. Frankrike har satt etikk i front av Frankrikes KI-strategi, og Tysklands Dataetiske kommisjonhar utviklet anbefalinger om hvordan man kan gå fra EUs etiske retningslinjer til et risikotilpasset regelverk.

  21. april i år presenterte EU-kommisjonen så sitt etterlengtede forslag til en forordning om harmoniserte regler for KI. Forordningen bruker en *risikobasert* tilnærming, og fokuserer på høyrisiko-anvendelser – definert som anvendelser som har potensielt negative effekter på helse, sikkerhet eller grunnleggende rettigheter for mennesker, som personvern eller diskriminering. Eksempler er KI-systemer som er innebygd i medisinsk utstyr, eller brukes til rekruttering, kredittvurdering eller vurdering av innbyggernes berettigelse til offentlig bistand.

  Høyrisiko KI-anvendelser vil måtte oppfylle visse krav, bl.a. en forsikring om tilstrekkelig høy kvalitet på data for trening, validering og testing av KI-systemer, og må også gjennomgå en samsvarsvurdering før de kan markedsføres i det europeiske markedet. Anvendelser som ansees å utgjøre en uakseptabel risiko blir forbudt. Eksempler på slike er beregning av sosiale

[16] https://www.cnbc.com/2021/02/19/australians-respond-to-facebooks-news-ban.html
[17] https://en.unesco.org/artificial-intelligence/ethics#recommendation
[18] EU-kommisjonens ekspertgruppe har skissert syv krav for at KI skal være pålitelig: (1) KI-baserte løsninger skal respektere menneskets selvbestemmelse og kontroll, (2) KI-baserte systemer skal være sikre og teknisk robuste, (3) KI skal ta hensyn til personvernet, (4) KI-baserte systemer må være gjennomsiktige, (5) KI-systemer skal legge til rette for inkludering, mangfold og likebehandling, (6) KI skal være nyttig for samfunn og miljø, og (7) Ansvarlighet. https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

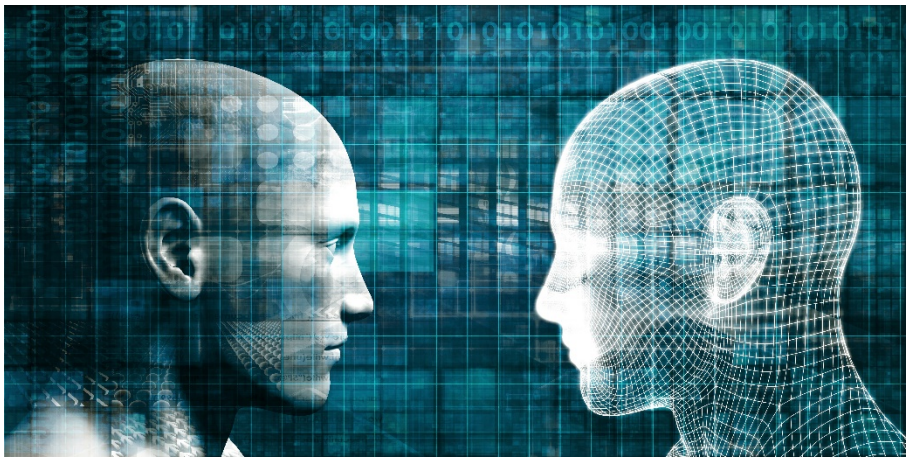
poengsummer (f.eks. til identifisering av utsatte barn som trenger sosial omsorg), utnyttelse av barn og psykisk funksjonshemmede (f.eks. gjennom virtuelle assistenter integrert i leker), og offentlig sanntids biometrisk identifikasjon (f.eks. ansiktsgjenkjenning) med formål om å håndheve loven (med noen unntak). Reguleringsforslaget vil nå gjennomgå en juridisk prosess i EU-systemet, og hvis det blir godkjent, tre i kraft innen 2–3 år.

- Andre land utenfor Europa – bl.a. Canada, Singapore, India– har også startet lignende aktiviteter.

Etiske og sosioøkonomiske hensyn knyttet til KI begynner også å dukke opp på radaren til verdens mest anerkjente KI-forskningsinstitutter, blant annet i Storbritannia, Canada og USA, og er dessuten blitt gjenstand for store grasrotkampanjer blant KI-forskere. Frivillige organisasjoner og forbrukerorganisasjoner øker også sin innsats for å fremme en mer menneskesentrert tilnærming til KI.

Verdens største internettgiganter – som Google og Microsoft– er dessuten blitt tydeligere på sin forpliktelse til å integrere etiske prinsipper for KI i sin virksomhet, og å maksimere de sosiale fordelene. Også europeiske selskaper (som f.eks. Telefonica, Deutsche Telekom, SAP) har kunngjort etiske KI-koder og styringsprinsipper.

Gitt trendene over, og et økende politisk og sosialt press, forventer vi at flere offentlige og private organisasjoner vil offentliggjøre etiske retningslinjer for KI som en del av sitt ansvarlige forretningsgrunnlag i årene som kommer. KI-etikk vil sannsynligvis snart påvirke mange selskapers forretningsmodeller i en eller annen form.



Figur 23.3 Etiske regler for teknologi som forretningskonsept. Illustrasjon: kentoh/Shutterstock.

Etisk KI i Norge: Hvordan kan vi bidra til den internasjonale debatten?

KI/ML er ennå i en tidlig utviklingsfase i Norge, og det er også den etiske debatten. Den norske regjeringen har fulgt EUs tilnærming til KI-etikk ved å integrere Kommisjonens etiske krav til pålitelig KI i Norges nasjonale KI-strategi.[19] Som en del av implementering av denne strategien åpnet

[19] https://www.regjeringen.no/no/aktuelt/regjeringen-legger-frem-nasjonal-strategi-for-kunstig-intelligens/id2685599/

Datatilsynet den første nasjonale *sandkassen for ansvarlig KI,*[20] med vekt på beskyttelse av personopplysninger i utviklingen og bruken av KI. Noen få store bransjeaktører, sammen med Negotia og en del offentlige organer, har erklært at de forplikter seg til utvikling av pålitelig, etisk og personvernbevarende KI. I det største NFR-finansierte forsknings- og innovasjonskonsortiet innen KI, SFI NorwAI, er en betydelig del av forskningen viet til pålitelig, forklarbar, personvernbevarende KI, også støttet av industripartnere som DNV GL, DNB, Schibsted, Telenor og andre. Flere fellesskap som fremmer høye dataetiske standarder dukker også opp, for eksempel blant informatikere og bedriftsledere.

Norges sosioøkonomiske modell er generelt basert på et høyt nivå av tillit og dialog mellom regjering, næringsliv og akademia. Dette gjelder også ny teknologiutvikling, herunder KI. Norge er kjent for høye standarder for tillit og demokrati. For et lite land som Norge er det fordeler ved å gå sammen med europeiske og andre vestlige allierte for å fremme den globale standarden for etisk KI. Med nasjonale fortrinn innenfor sterke globale næringer (som maritim, olje, fiskeoppdrett), der bruken av data og KI vil definere fremtidig konkurransekraft, ligger det enorme muligheter for Norge i å utvikle KI for sosialt gode produkter med høy etisk standard.

Debatten om KI-etikk er som vi har sett, en mangesidig debatt, som ofte starter med «det kommer an på» – hvilke moralske verdier en person bekjenner seg til, hvilke kulturer hun/han representerer, hvor hun/han er i rom og tid m.m. Ulike perspektiver er verdifulle, og det er viktig å både lære om og respektere forskjellige synspunkter. Arbeidet med å skape en global standard for etiske KI-prinsipper og verdier bør ta inn over seg dette, samtidig som det baseres på rettssikkerhet, demokrati, og individers og samfunnets velferd.

[20] https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/