ane | ASSOCIATION OF NORDIC ENGINEERS

NITO

Etikk og kunstig intelligens

- Rapport fra hackaton 25. september 2018

IT UNIVERSITY OF COPENHAGEN

ETHOSLAB

Innhold

Innledning

Ideen om eksistensen av kunstig intelligens har eksistert i flere århundrer, men samordnet forskning på KI begynte ikke før i kjølvannet av 2. verdenskrig. Dette skjedde i USA, og var inspirert av krigens teknologiske nyvinninger innen signaldeteksjon, kodeknekking og ammunisjonssporing. På dette tidspunktet var målet å utvikle enheter som kunne handle intelligent og autonomt gjennom forene vitenskap og teknologi. og. Det første nevrale nettverket ble bygget så tidlig som på 1950-tallet av en gruppe forskere ved Massachusetts Institute of Technology (MIT). Målet til forskningsmiljøet i denne tidlige fasen viste seg å være i overkant optimistisk og innen få tiår kretset debatten rundt KI igjen rundt spådommer – samtidig som arbeidet med de underliggende tekniske prosessene og innovasjonene fortsatte med full styrke.[1] Uroen som har spredt om seg i det siste kan gi inntrykk av å være av nyere dato og kun gå tilbake til de to første to tiårene av det 21. århundre. I realiteten er det en stor grad av kontinuitet i den teknologiske utviklingen vi i dag kjenner under betegnelsen KI. Eksempelvis ble det i løpet av de siste to tiårene av det 20. århundre utviklet ekspertsystemer for å levere tekniske informasjonsprosesser for å støtte menneskelig beslutningstaking. I forbindelse med implementeringen av disse systemene ble det uttrykt bekymring for at fordommer kunne innarbeides i disse. Tidlige anbefalinger gikk derfor ut på at disse systemene kun skulle brukes av de menneskelige beslutningstakerne som et rådgivende verktøy.[2] Selv om mye av den aktuelle diskusjonen rundt KI fortsatt har et futuristisk preg, er sannheten at vi – i det minste i den vestlige verden – allerede har rullet ut mange autonome systemer og at disse oppfyller en rekke samfunnsfunksjoner, bla. i sykehus, statlig forvaltning og innen beslutningstaking på ledelsesnivå.[3] KI gjorde sitt inntog i private hjem lenge før SIRI, selv om SIRI kanskje er det første eksempelet som er spesielt kommunikativ.

Framskritt innen prosessorhastighet, vekst i omfang og kvalitet av ulike typer nettverk, og framvekst av stordata er sentrale drivere av et stadig høyere tempo innen forskning og nyskapinger innen maskinell læring, datautvinning og nevrale nettverksapplikasjoner. Uroen rundt denne utviklingen øker dermed også kraftig. Aktuelle diskusjoner rundt KI-relaterte problemer må derfor rette seg mot langt mer kraftige og komplekse teknologier enn det som ble diskutert for relativt få år siden. Dette til tross for at bekymringene som er uttrykt er nokså like. Hvordan kan vi utvikle KI-teknologi som har en positiv effekt på samfunnet?

Fremveksten av utallige KI-systemer og -løsninger har hatt stor innvirkning på hvordan man forstår og møter prediksjon. Der bruken av datasystemer til prediksjon tidligere utgjorde en betydelig økonomisk investering er dette nå blitt en selvfølgelig del av enhver datamaskin. Etterhvert som prediksjon er blitt alminneliggjort, gjenstår spørsmålet om hvem som kan vurdere hvor riktige disse prediksjonene er? På tross av dagens satsing på automatisering av mange prosesser og praksiser, er det mange som påpeker at KI den dag i dag er «laget av mennesker»[4] ettersom også KI fortsatt bygger på menneskelig skjønn når det gjelder dataene som skal mates inn i systemet - – samt i prosessens sluttfase.[5] Ettersom prediksjon og eksperimentering er uløselig knyttet opp mot hverandre, åpner innføringen av prediksjon i nye områder også opp helt nye muligheter for eksperimentering. Etter hvert som eksperimentering spiller en rolle i langt flere bruksområder på tvers av autonome systemer oppstår også nye etiske problemstillinger.[6] Hvordan kan ingeniører utvikle systemer som er «beviselige trygge» også etter såkaltrekursive selvlæringsprosesser? Er det behov for å utvikle en ny tilnærming til hvordan ingeniører arbeider med sikkerhet?[7]

teknologien I tillegg til utfordringen med fordommer i databehandling er det også en utfordring med mangfold i arbeidsstyrken som både utvikler og implementerer KI- Mange slår fast at den iøynefallende mangelen på mangfold i de tekniske yrkene er et vesentlig problem og at dette kan være en av hemskoene som sinker arbeidet med å håndtere fordommer i algoritmiske systemer.

«Det å fokusere på menneskelige beslutningers rolle i å skape teknologier er en måte å beholde ansvaret og ivareta hverandre – for 'teknologier bryr seg ikke'».[8]

Slike analyser av ansvarsforhold er nødvendige når man møter teknologiske endringer som omkonfigurerer maktforholdene i vår samfunnsstruktur. I dette landskapet blir eksisterende etiske regler et upålitelig kompass og vår evne til å forutsi de potensielle konsekvensene av design og implementering blir begrenset.

For å vurdere disse spørsmålene innenifra og sammen med fagfolk, og for å styrke menneskers rolle i utviklingen av KI, avholdt ANE i samarbeid med IT-universitetet i København 25. september 2018 en etikk-hackaton med navn «Nordiske ingeniørers standpunkt om EUs fremtidige rammeverk for KI og etikk». Her samlet vi IKT-ingeniører fra de fem nordiske landene for å skape et felles standpunkt tuftet på praktisk erfaring og aktuelle debatter rundt KI og etikk. Strategidokumentet du holder i dine hender, med sine anbefalinger og retningslinjer, er et resultat av hackatonen. Dokumentet gjenspeiler derfor nordiske ingeniørers omforente standpunkt på KI og etikk.

RAPPORTENS STRUKTUR

Del 1 (ANEs hackaton) beskriver selve hackatonen, hvordan den var strukturert og fremgangsmåten under hackatonen.

Del 2 (Operasjonelle definisjoner) inneholder en diskusjon om hvordan medlemmene i ANE er blitt enige om å definere de operasjonelle begrepene rundt KI og etikk i sine diskusjoner.

Del 3 (Viktige etiske spørsmål i forbindelse med KI) er en oversikt over de mest fremtredende problemstillingene og bekymringene man må gripe fatt i for å oppnå etiske måter å arbeide med KI på. Disse spørsmålene er transparens, klare ansvarsforhold og tillit, unngåelse av skade og fordomshåndtering.

Del 4 (Arbeid med de viktige spørsmålene) vurderer hvilke tiltak som kan iverksettes i forbindelse med spørsmålene som er tatt opp i den forrige delen og hva dette betyr konkret.

Del 5 (Anbefalinger og retningslinjer) beskriver anbefalinger for arbeidet med etiske spørsmål relatert til KI for den enkelte ingeniør, ingeniørfaglige organisasjoner og myndighetene.

ANEs HACKATON

25. september 2018 arrangert ANE, sammen med IT-universitetet i København, en etikk-hackaton med navn «Nordiske ingeniørers standpunkt om EUs fremtidige rammeverk for KI og etikk». Målet med workshopen var å utvikle et sett med anbefalinger og retningslinjer som alle har bidratt til denne rapporten.

Workshopen samlet en gruppe ingeniører med ulik bakgrunn fra de fem nordiske landene. Blant deltakerne var medlemmer fra nordiske fagorganisasjoner: Sveriges Ingenjörer (Swedish Association of Graduate Engineers), Norges Ingeniør- og Teknologorganisasjon (NITO), Verkfræðingafélag Íslands (Association of Chartered Engineers in Iceland, kjent som VFI), Dansk ingeniørforening (Danish Society of Engineers, kjent som IDA) og finske Teknikens Akademikerförbund (Association of Academic Engineers and Architects, kjent som TEK). De fleste av deltakerne var aktive som IT-ingeniører. Et mindretall arbeidet innen faglig relevant forskning.

Workshopen var bygget opp rundt et rammedokument med en oversikt over aktuelle debatter rundt KI og etikk. Rammedokumentet var forfattet av forskere ved IT-universitetet i København som hadde gjennomgått forskningslitteraturen og tidligere utgitte etiske retningslinjer for KI. Før workshopen ble rammedokumentet sendt ut til deltakerne.

Aktivitetene på workshopen var basert på debattene som ble fremlagt i dette dokumentet. Deltakerne ble delt inn i fem grupper og hver gruppe ble bedt om å arbeide med én problemstilling i dokumentet i løpet av dagen. Først ble alle gruppene bedt om å komme med en arbeidsdefinisjon av KI og etikk. Øvelsen besto av et selvstendig element, der deltakerne ble bedt om å reflektere rundt sin egen yrkeserfaring, og en gruppeaktivitet, hvor deltakerne utarbeidet en definisjon som gjenspeilte gruppens omforente standpunkt ut fra sine selvstendige refleksjoner. Avslutningsvis ble to av gruppene bedt om å presentere sine definisjoner i plenum. Det ble så åpnet for at alle deltakerne kunne kommentere fremleggene.

I den andre oppgaven fikk gruppene tildelt en av fem ulike problemstillinger fra rammedokumentet og bedt om å diskutere disse ut fra sine egne faglige erfaringer. Med utgangspunkt i den tildelte problemstillingen, fant deltakerne fram til eksempler på denne fra sin praksis innen utvikling av KI og diskuterte betydningen av disse. Denne øvelsen ble etterfulgt av fremlegg ved to grupper og en plenumsdiskusjon.

I den siste oppgaven ble deltakerne bedt om å benytte det de var kommet frem til under de to foregående oppgavene som et springbrett for å foreslå praktiske retningslinjer som kunne brukes for å utarbeide dette strategidokumentet.

Deltakerne diskuterte etiske måter å arbeide med KI og vurderte hva de selv ønsker seg i forbindelse med utvikling eller implementering av KI, eller hva de ønsker å dele med yngre kolleger. Hver gruppe presenterte sine forslag i plenum som en del av en grundig avrundende diskusjon. Disse retningslinjene og anbefalingene utgjør kjernen i ANEs synspunkt på KI og etikk og gjenspeiles i dette dokumentet.

Operasjonelle definisjoner

Når man utformer et rammeverk ut fra felles samtaler er det viktig å etablere omforente definisjoner av de viktigste begrepene. I dette tilfelle var det behov for å definere begrepene KI og etikk. Hackaton deltakerne ble utfordret til å diskutere sine egne definisjoner av disse to begrepene ut fra rammedokumentet og så komme frem til en omforent forståelse. I avsnittene nedenfor foreligger definisjonene som opprinnelig ble utarbeidet ut fra den eksisterende litteraturen og så videreutviklet gjennom samtalene med ANEs medlemmer.

DEFINISJON AV KUNSTIG INTELLIGENS (KI)

Fra Roomba-støvsugere til Siri og andre mobilapplikasjoner omringes vi i økende grad av systemer som ikke bare forstår at man henvender seg til dem, men også svarer på hensiktsmessige måter. Selv om disse systemene ikke er «intelligente» i ordets rette forstand, fungerer de godt i den konkrete sammenhengen. Disse systemene – fra militære droner og lagerroboter til GPS-systemer for bil og robotassistenter for eldre – er eksempler på den stadig økende bredden i bruksområder for KI. Selv om KI har overgått mennesker i mange konkrete områder – slik som sjakk – hersker det nærmest fullstendig enighet blant KI-eksperter om at KI ikke kan måle seg med mennesket når det gjelder kritisk evne.[9] Denne forståelsen av menneskelige evner og hvilke menneskelige egenskaper maskinene bør etterligne kommer tydelig frem i hvordan de som arbeider med slike systemer definerer KI.

Det finnes per i dag ingen universell definisjon av KI. Men flere definisjoner er godt etablerte. EU-kommisjonens uttalelse om Kunstig Intelligens i Europa benytter følgende definisjon:

«Kunstig Intelligens (KI) handler om systemer som utviser intelligent atferd ved å analysere sine omgivelser og handle – med en viss grad av autonomi – for å oppnå konkrete mål.»[10]

Ifølge EU-kommisjonen[11], har noen definisjoner av KI fokus på gjenstanders autonomi, mens andre har fokus på KI som en samling raskt sammensmeltende, smarte digitale teknologier som ofte er sammenhengende, forbundet eller fullstendig integrert. Sistnevnte gruppe inkluderer klassisk KI, algoritmer for maskinlæring, dyp læring og konneksjonistiske nettverk, GAN-nettverk (GAN står for *generative adversarial networks*), mekatronikk og robotikk. Vi ser hvordan disse teknologiene har smeltet sammen i nyvinninger som chatte-roboter, robotiserte våpensystemer, tale- og billedgjenkjenningssystemer og selvkjørende biler.

Ifølge uttalelsen fra Institute of Electrical and Electronics Engineers (IEEE) om «Ethically Aligned Design», er «KI (vanligvis digitale) gjenstander som kombinerer hvilken som helst av følgende egenskaper: evnen til å oppfatte handlingskontekster, evnen til å handle og evnen til å knytte kontekster til handling.» Videre skriver IEEE i sin uttalelse: «Etter hvert som bruken og effekten av autonome og intelligente systemer (K/IS) blir utbredt er det et behov for å etablere samfunnsmessige og strategiske retningslinjer slik at disse systemene forblir menneskesentrerte og støtter opp om menneskehetens verdier og etiske prinsipper. Disse systemene må oppføre seg på en måte som er til menneskers fordel utover det å oppnå funksjonelle mål og håndtere tekniske problemer.»[12] Denne definisjonen av KI er svært bred. Resonnementet om at evnen til å handle og at selve handlingen må være i tråd med menneskehetens verdier og prinsipper søker å avgrense typen aktiviteter det her er snakk om.

Store teknologikonsern med store investeringer i KI har også supplert med egne definisjoner. Google har eksempelvis bidratt med en tilsynelatende enkel definisjon: «Essensen i KI er

dataprogrammering som lærer og tilpasser seg.»[13] IBM lar være å definere begrepet KI. I stedet peker selskapet på mangfoldiggjøringen av det IBM beskriver som «K*-algoritmen» som «et sentralt verktøy for KI og som man finner i alle lærebøker om KI.»[14] Fremfor å kalle bredden av de stadig mer autonome systemene i verden «KI», er IBM opptatt av spredningen i bruken av KI-teknologier innen systemdesign – det være seg maskinlæring, dyp læring eller GAN-nettverk.

Under hackatonen ble det uttrykt en rekke bekymringer blant ANEs medlemmer rundt bruken av begrepet kunstig intelligens. De fleste av deltakerne ønsket ikke å bruke ordet intelligens fordi definisjonen av ordet er såpass kompleks. En av deltakerne forklarte dette som følger: «Det er ikke engang klart hva intelligens er eller hva slags nivå som kreves.» Gitt at begrepet i senere tid igjen har fått en såpass dominerende plass, ble gruppen likevel enig om at det ikke var hensiktsmessig å omdefinere begrepet. Noen foreslo at: «På en måte kan KI defineres som et system som kombinerer automatisert automatisering og maskinlæring med en generell forståelse av kontekstuell bevissthet og tilpasningsevne». Deltakerne var enige om at uansett hvilken definisjon man bruker, er ikke KI én form for teknologi, men en gruppe teknologier som har til felles at de demonstrerer en form for bevissthet, autonomi og tilpasningsevne mht. oppgave- og prosessautomatisering. Fremfor å definere hva som regnes som «intelligent» og hva som ikke gjør det, foreslo ingeniørene fra ANE i stedet at det ikke finnes noe klart skille mellom KI og andre dataprogrammer. Deltakerne ser på KI som et spekter, et bredt utvalg metoder som fører til ulike teknologier som defineres ut fra nivået på tilpasningsevne og autonomi. Historien har tross alt vist at det vi anser for å være KI endrer seg over tid.

«Mange ingeniører foretrakk begrepet 'maskinell intelligens' eller til og med 'utvidet intelligens', da dette er begreper som formidler at teknologien det er snakk om er et sett med verktøy fremfor et eget system. Problemstillingen er ikke mennesket kontra maskinen, men hvordan mennesker skal implementere måten maskiner tenker på.»[15]

ETIKK

Etikk er blitt et moteord når man diskuterer teknologi generelt, og når man diskuterer KI konkret. Mediene har artikler der man debatterer etikk, store bedrifter satser på å etablere etiske utvalg og ber organisasjoner fra sivilsamfunnet om å gjennomføre ulike evalueringer og forskningsprosjekter. Den aktive diskursen rundt teknologietikk er blitt kritisert for å være en måte næringsinteressenter innen teknologi kan omgå regulering – nettopp ved å henvise til etikk som en slags myk regulering og en måte å synliggjøre positive verdier overfor samfunnet.[16] Spørsmålet er selvfølgelig ikke hvorvidt man skal handle etisk, men hva etisk atferd betyr i forbindelse med KI-teknologier.

Nyere tekster om etiske utfordringer mht. teknologi viser til et bredt utvalg etiske rammeverk. Jevnt over kan man kategorisere disse bekymringene i to overordnede grupperinger: konsekvensetikk og utilitaristisk etikk. Mye av den etiske vurderingen av de fremvoksende teknologiene dreier seg om spørsmålet om hva som er positivt og hva som er negativt med produktene, tjenestene og prosessene de kan føre til, og hva som er rett eller galt med måten de kan brukes på.[17] Noen vurderinger som f.eks. diskusjoner rundt selvkjørende biler, har hatt et konkret fokus: man har sett på det utilitaristiske ønsket å minimere skadevirkninger og maksimere fordeler for alle berørte parter. Dilemmaet i disse vurderingene er hvordan man skal definere skade kontra fordeler og hvordan identifisere hvem som bør inkluderes i slike kalkyler.[18]

På generelt nivå gjelder etikk de rammene og prinsippene som definerer enkeltpersoners evne til å leve gode liv og som klart fastsetter enkeltpersoners rettigheter, forpliktelser og ansvar.

NITO uttrykker dette slik: «Etikken gir oss verken anbefalinger eller pålegg. Den gir oss i stedet praktiske verktøy til å skille mellom gode og dårlige begrunnelser, og dermed kunne ta kloke beslutninger.»[19] Selv om definisjonen er lite presis, inneholder mange av de etiske retningslinjene nevnt ovenfor praktiske anbefalinger og tar for seg generelle prinsipper. Noen av disse diskuteres nedenfor.

Den britiske standarden BS8611:2016, med navn «Robots and Robotic Devices: Guide to the Ethical Design and Application of Robots and Robotic Systems», definerer etikk simpelthen som «en felles forståelse av prinsipper som avgrenser og er veiledende for menneskelig atferd.»[20] IEEEs «Statement on Ethically Aligned Design», på den andre siden, definerer ikke etikk. Uttalelsen fastslår kun at autonome og intelligente systemer «må oppføre seg på en måte som er fordelaktig for mennesker utover det å oppnå funksjonelle mål og håndtere tekniske problemer.»[21]

«Debatten om etikk er diskusjoner om de forpliktelsene som ingeniører må påta seg i sitt virke og i sin praksis. Etiske forpliktelser har to aspekter: yrkesetiske og personlige forpliktelser.[22] Førstnevnte etablerer retningslinjer for beslutningstaking og atferd innen faglig praksis. Sistnevnte sikrer at enkeltpersoner reflekterer og handler i de situasjoner der de faglige retningslinjene ikke strekker til.»

Yrkesetikk skisserer hvordan bredere etiske normer – slik som ansvar, integritet, rettferdighet, transparens og skadeunngåelse – gjelder for det arbeidet som ingeniører utfører i praksis. Det å være en yrkesutøver betyr å være en del av et moralsk fellesskap sammen med andre som har del i det samme ansvaret og å kunne dra nytte av andres erfaringer for å kunne handle når man møter lignende moralske dilemmaer, vanskelige avgjørelser eller negative konsekvenser. Det personlige aspektet betyr at enkeltpersoner ikke er likegyldige mht. hva slags innvirkning de har på andres liv i de situasjonene der de yrkesetiske retningslinjene ikke strekker til. Personlig etikk setter ingeniører i stand til å ta ansvar for sine egne moralske valg og konsekvensene av disse i de tilfeller arbeidsgivernes moralske valg ikke er sammenfallende med den enkelte ingeniørs etikk.

Selv om faglige veiledere og etiske retningslinjer inneholder ulike anbefalinger, har de ikke til hensikt å fungere som sjekklister eller utfyllende rettledninger om etisk atferd i enhver tenkelig situasjon som ingeniører kan støte på. Disse veilederne og retningslinjene skal kun være verktøy for å hjelpe ingeniører vurdere hva som er «passende» i en gitt situasjon. Med utgangspunkt i dette har medlemmene i ANE utviklet sine egne definisjoner og stridsregler.

Gjennom hele hackatonen erkjente ANEs medlemmer at etikk også er et spørsmål om kulturelle verdier og ikke nødvendigvis uforanderlig over tid. Deltakerne stilte spørsmål om hvorvidt KI-relatert etikk overhodet måtte skille seg fra annen etikk og tok utgangspunkt i eksisterende retningslinjer. Man ar enig om at man kan se på etikk som veiledende prinsipper for menneskelig atferd. En av deltakerne kommenterte: «Det er vanskelig å handle rett – på samme måte som det er vanskelig å lage de rette tingene. Etikk spiller inn gjennom hele prosessen.» Deltakerne var enige om at det var et behov for etiske prinsipper i forbindelse med utviklingen av autonome systemer – om ikke annet som rammer og retningslinjer, avhengig av hva som vurderes som moralsk, kulturelt og sosialt akseptabelt. Diskusjonen fokuserte mer på hvem som får bestemme hva som er akseptabelt og under hvilke forhold.

Mens noen moralske prinsipper ikke er universelle og kan endres over tid, finnes det noen som er bestandige, slik som prinsippene i FNs menneskerettighetserklæring. En nyere rapport fra tankesmien «Data & Society» argumenterer at Menneskerettighetserklæringen kan være et nyttig hjelpemiddel for å styre KI-utviklingen.[23] Erklæringens krav om respekt for menneskeverd, vernet av ikke-diskriminering og likeverd samt uttrykksfrihet er idealer som alle kan være enige i rent prinsipielt. Utfordringen ligger i hvordan slike krav kan overføres til praktisk beslutningstaking. Deltakerne på hackatonen var opptatte av hvem som skal holdes ansvarlig for overholdelse eller brudd på disse prinsippene innen KI-design. Det er sant at teknologiselskaper og organisasjoner er svært opptatte av å sørge for etterlevelse av eksisterende personvernlovgivning og annet virksomhetsrelevant regelverk. Men i og med at KI-teknologier utvikles i rasende fart er det ikke sikkert at det holder å følge regelverket. Hvem skal få oppgaven med å definere grensene og retningslinjene som går utover det lovmessige i det enkelte prosjektet eller den enkelte organisasjonen? Det finnes ikke noe enkelt svar på spørsmålet om hvor ansvaret ligger eller hvem som skal holdes ansvarlig for eventuelle negative følger. Uansett hvor prinsippfaste eller opptatt av etisk atferd ingeniørene er, er det ikke sikkert at dette er nok i store distribuerte prosjekter der den overordnede strukturen kan være vanskelig å tyde for personer som arbeider med forskjellige sider av et system. Det er opplagt at etisk debatt ikke kun er den enkelte ingeniørs ansvar. Deltakerne i hackatonen var veldig tydelige på at etiske diskusjoner rundt teknologien som nå utvikles og bygges må skje på en rekke ulike nivåer innad i organisasjonene.

Viktige etiske spørsmål i forbindelse med kunstig intelligens

Forventningene til KI øker stadig. Dette kommer tydelig fram i tre omfattende rapporter fra 2016, utgitt av henholdsvis den amerikanske regjeringens Kontor for vitenskap- og teknologistrategi (Office of Science and Technology Policy, forkortet som OSTP), det britiske Underhusets vitenskaps- og teknologikomité (Science and Technology Committee) og Europaparlamentets Utvalg for rettslige spørsmål (Committee on Legal Affairs). I disse rapportene finner man en visjon for hva som skal til for å forberede oss til en fremtid der KI vil være implementert i alle sider av vårt moderne samfunn. Cath m.fl.[24] sammenlikner disse tre rapportene og deres respektive syn på det grunnleggende spørsmålet rundt KI: hva blir KIs etiske, sosiale, økonomiske og politiske effekt? Hver rapport presenterer sin egen visjon for «et godt KI-samfunn»[25] og hva slags lovgivende rolle hver statsmyndighet er villig til å påta seg, ut fra sine ulike tradisjoner for statlig styring. Mens det amerikanske OSTP ser for seg at teknologisektoren vil regulere seg selv, tilrår EU-rapporten utviklingen av nye institusjonelle ordninger og rettslige strukturer for å imøtegå mulige risikoer, samtidig som man støtter forskning og utvikling.

Felles for alle rapportene er at de maner til økt forskning og utvikling innen KI for å kunne dra fordeler av det iboende potensialet i KI. Samtidig påpeker man at det må gjøres en innsats for å sikre at utviklingen av disse teknologiene sikrer transparens, klare ansvarsforhold og at teknologien utformes i tråd med menneskets verdier. Det er stort fokus både på å minimere fordommer i KI-systemene som utvikles og på å sikre mangfold i arbeidsstyrken. Videre tar rapportene for seg at utdanningssystemene kanskje bør legges om for å sette dem i stand til å møte de stadig voksende behovene og bekymringene.[26] Utfordringene rapportene ser for seg i forbindelse med bred implementering av KI-systemer har mange paralleller med det som diskuteres i forskningsmiljøene, mediene og i forbindelse med politiske beslutningsprosesser.

I denne delen presenterer vi problemstillingene som tas opp i eksisterende og nyutviklede internasjonale etiske retningslinjer og bekymringer som er formidlet vedrørende etikk i forbindelse med KI, programvare eller digitale teknologier generelt (vedlegget inneholder en liste med relevante rapporter) samt det som ble diskutert av deltakerne på ANEs hackaton. Problemstillingene er samlet under fem kategorier som til dels overlapper med hverandre: samfunnsansvar, transparens, klare ansvarsforhold og tillit, unngåelse av skade og hvordan imøtegå fordommer.

SAMFUNNSANSVAR

Ingeniører er en yrkesgruppe med lange tradisjoner når det gjelder diskusjoner om plikt og ansvar. Ved århundreskiftet innførte canadiske ingeniører en seremoni kjent som «Kallsritualet for ingeniører». Her fikk unge ingeniører en ring laget av jern som de skulle bruke på lillefingeren gjennom hele sin profesjonelle karriere. Ringen skulle være en påminnelse om deres plikter og ansvar.[27] Ringen handler ikke om å markere ingeniørenes kvalifikasjoner, men er av stor symbolsk betydning. Den understreker hvilken makt ingeniører har og at denne må utøves med tanke på fremme positive verdier. Som yrkesutøvere har ANEs medlemmer – i kraft av sin sentrale rolle innen design, utvikling og produksjon av teknologier – et ansvar overfor samfunnet. ANEs medlemmer er opptatt av ulike teknologiers rolle i samfunnet og hvordan de bidrar til å strukturere samfunnet.

Med utgangspunkt i spørsmålet «Hvilke verdier ønsker vi å organisere våre samfunn rundt?» diskuterte deltakerne på workshopen hvordan teknologier bidrar til å forme samfunnsstrukturer, spesielt teknologier som benytter KI-metoder. Ideen om å minimere teknologienes negative konsekvenser er en felles forutsetning de fleste kan enes om, men i realiteten er den altfor generell til å være praktisk omsettbar. Ved å spisse spørsmålene kan vi finne fram til mer konkrete svar rundt

samfunnsansvar. Eksempelvis kan man spørre «Hvem tjener på utviklingen av KI? Er dette kun noen få personer, bestemte grupper eller en større gruppe?»

Ingeniørene som deltok i hackatonen vurderte ulike ansvarsforhold, fra å sørge for positive innvirkninger på samfunnet til å verne om demokratiske prosesser. Eksempelvis diskuterte ingeniørene problemet med innblandingen i det amerikanske presidentvalget i 2016 og ingeniørenes ansvar for å bygge systemer som kan forebygge slikt. Samtidig var deltakerne klar over begrensningene både forbundet med måten konsepter utvikles på, og forestillinger forbundet med roller og makt. Mange påpekte at det var ulik tilgang til maktutøvelse i organisasjoner hvor ingeniører ofte ikke har samme beslutningsmyndighet som deres ledere. Det ble også nevnt at det kunne være vanskelig å forutse de utilsiktede konsekvensene av IT-ingeniørers arbeid. En deltaker nevnte eksempelet Airbnb: ideen om å leie ut rom til en billig penge er god, men i virkeligheten har nettsiden ført til at leilighetspriser i populære byer har økt slik at lokalbefolkningen ikke lenger har råd til å leie eller kjøpe bolig. Systemer kan med andre ord misbrukes. Dette er spesielt problematisk sett opp mot hvordan penger og makt samles og fordeles, noe som igjen kan føre til at innovasjon stopper opp. Det haster derfor med å finne ut hvordan problemer kan avdekkes og hvordan man kan finne ut hvorvidt foreslåtte løsninger igjen kan avstedkomme nye problemer. Flere av deltakerne på hackatonen sa at det sentrale spørsmålet var: Hva optimaliserer vi og hvem optimaliserer vi det for? I en situasjon der det vies mye oppmerksomhet til optimalisering innen teknisk utvikling og innovasjon, står dette spørsmålet sentralt. Selv om tanken på å vurdere teknologienes samfunnsmessige effekt kan virke overveldende, er det helt vesentlig å vite hvordan man kan stille de rette spørsmålene og å forsøke å forstå teknologi i et større perspektiv.

TRANSPARENS

KI-teknologier er ikke lette å forstå for de som ikke arbeider med å designe og utvikle denne typen teknologi. Fordi KI-systemenes algoritmer er ugjennomtrengelige kan til og med mennesker med mye kunnskap om temaet synes at det er vanskelig å forstå hvordan KI-programvare og -utstyr produserer utdata. Denne ugjennomtrengeligheten gjør det vanskelig å vite hvordan beslutninger ble tatt, hvorvidt det er gjort feil og hvordan disse eventuelt kan ha oppstått. Det kan derfor være svært utfordrende å forklare den underliggende logikken til et konkret system til en større gruppe, det være seg andre yrkesutøvere eller representanter for allmennheten som berøres av systemet. Etterhvert som systemer lærer å utføre oppgaver på mer og mer autonomt vis, dvs. uten tilsyn fra mennesker, kan de produsere utfall som ikke var forutsett av menneskene som sto for den opprinnelige utformingen. Mange har foreslått å takle dette ved å sikre åpenhet rundt hvordan autonome systemer fungerer overfor alle aktuelle interessenter. IEEEs «Vision for Ethically Aligned Design»[28] påpeker at «begrepet transparens også omfatter begrepene sporbarhet, forklarbarhet og tolkbarhet.» For mange av de eksisterende tekstene som tar for seg etiske spørsmål rundt KI, er transparens også nøkkelen til samtykke. Informert samtykke kan ikke gis uten at personen det gjelder forstår implikasjonene. Spørsmålet er derfor: Hvordan oppheve motsetningsforholdet mellom den høyst innviklede indre funksjonaliteten til systemene som utvikles av ingeniørene og behovet for å forklare hvordan disse fungerer overfor utenforstående?

Transparens blir ofte nevnt som en løsning på mange av de etiske utfordringene rundt KI-systemer og deres funksjon, men transparens er ikke en fullgod løsning. Faktisk har transparens en rekke begrensninger, gitt at det å ha transparente prosesser ikke er ensbetydende med å være forståelig eller at man kan handle ut fra de opplysningene som legges frem. Videre finnes det omstendigheter

der fullstendig transparens kan være til skade.[29] Selv om det først og fremst er selskaper som fremhever behovet for å beskytte forretningshemmeligheter som et argument mot transparens, er det viktig å nøye vurdere spørsmålene om hva som skal synliggjøres, og overfor hvem og hvorfor. I noen tilfeller kan satsing på transparens føre til at det genereres så mye informasjon at det viktige kan drukne i et hav av opplysninger uten at dette er intensjonen. Hvor mye som skal legges frem, når og overfor hvem er alt annet enn innlysende, gitt kompleksiteten i KI-systemer. Det er også viktig å huske at forsøk på å være transparent ikke nødvendigvis er tillitsskapende.[30] Alt dette betyr at transparens – på tross av å være en positiv målsetting – krever at man tar høyde for en rekke potensielle fallgruver.

Deltakerne på workshopen var opptatt av å diskutere transparens og vurderte spørsmålene rundt dette nøye. En deltaker tok opp en ofte omtalt utfordring: algoritmiske beslutningsprosesser og deres ugjennomsiktighet. Deltakeren påpekte at kravet om transparens også retter oppmerksomheten mot mangler ved dagens beslutningstakingsprosesser: heller ikke menneskelig beslutningstaking er transparent. Tvert imot er prosessen bak menneskers beslutningers ofte utydelig. Et eksempel er dommeres avgjørelser i forbindelse med prøveløslatelser. Det viser seg at slike avgjørelser kan påvirkes av når på dagen de fattes. En rekke potensielle løsninger ble foreslått: noen sa at det er mulig at fremtidens design vil måtte inkludere elementer som «Innebygd transparens» i alle trinn av utviklingsprosessen. Andre påpekte at det finnes et behov for uavhengig verifisering og at denne tilnærmingen kan være en måte å møte problemet på.

Deltakerne på ANEs hackaton anerkjente også at det å satse på transparens i utformingen av KI-teknologier kun er en del av løsningen. Det er også viktig med transparente beslutningsprosesser gjennom hele organisasjonen i enheter som utvikler slike teknologier. Dette må synliggjøres både internt og eksternt for å fremme klare ansvarsforhold og fostre økt tillit.

ANSVARSFORHOLD OG TILLIT

I all teknologisk utvikling er spørsmål om ansvarsforhold og tillit uløselig forbundet med strukturene i organisasjonene som produserer bestemte teknologier. Dette gjelder uansett om det er snakk om en statlig bygd bro eller en plattform for sosiale medier utviklet av et privat konsern. Tillit fra de som berøres av disse teknologiene henger sammen med at det etableres ansvarsledd både innad i og utenfor organisasjonene som er ansvarlige for selve teknologien. Selv om transparens kan fremme klare ansvarsforhold, trenger dette ikke nødvendigvis føre til tillit til slike teknologier.[31]

Uro rundt ansvarsforhold og tillit i forhold til autonome systemer går igjen som en rød tråd i dagens diskusjoner om etikk og KI, og nevnes bl.a. i EUs uttalelse om KI, robotikk og «autonome» systemer, i IEEEs visjon om «Ethically Aligned Design», og i alle tre offisielle plattformer for fremtidens KI[32] – for å nevne noen. Her hersket det stor enighet om at ingeniører som opplever at KI-systemer handler på en måte man ikke har forutsett eller forstått ikke kan løsrive seg fra sitt etiske ansvar ved å vise til uvitenhet. Det er klart at designere og utviklere må stå til ansvar for følgene av eget arbeid. Men hvor går skjæringspunktet for dette ansvaret og målsettingene til organisasjonene som lager disse teknologiene? For autonome systemer gjelder spørsmålet: «hvem er ansvarlig når ting gjør ting?»[33]

Hackaton deltakerne gjentok en rekke ganger at «klare ansvarsforhold går hånd i hånd med transparens.» Likevel er det slik at «det å oppnå tillit og sikre klare ansvarsforhold ikke kan skilles fra hverandre – de to henger sammen» og at tillit «raskt kan gå tapt dersom den misbrukes». På tross av

fantasifulle spådommer om nye KI-systemer på full fart inn i vår verden, er virkeligheten at mange av dagens KI-systemer er langt fra perfekte. De er ofte utdaterte og kan også være uforutsigbare. I en verden der kravene til KI ofte defineres av kunder eller selskaper, hvordan kan den enkelte ingeniør utøve innflytelse på spørsmål om ansvarsforhold eller tillit? Man vil alltid måtte ta ansvar for personlige handlinger og for faglig forsvarlig atferd, men hvor langt kan man strekke en persons ansvar for autonome systemers handlinger? Noen ingeniører foreslo at det er viktig å etablere organisasjonsinterne prosesser som sikrer god ansvarsfordeling. En løsning som ble foreslått var å etablere «kontrollstasjoner» på ulike stadier av applikasjoners livssyklus (planlegging, implementering, produksjonssetting). På denne måten kan ingeniører og/eller brukere ta det ansvaret for de eventualitetene som svarer til det aktuelle utviklingstrinnet. Det viktige her er at ansvar ikke kun er designerens eller ingeniørens, men noe som må fordeles på tvers av hele prosessen og blant prosessens ulike interessenter. Det er en forutsetning at alle interessenter utvikler en bevissthet om potensielle problemstillinger, herunder fremtidig bruk av verktøy for risikovurdering og verifisering.

«Mange KI-ingeniører insisterte på at arbeidet med KI måtte følge den gylne regelen når man tenker på framtidige brukere. Man må huske at utviklerne har et mye mer omfattende beslutningsgrunnlag enn de fleste interessenter.»

UNNGÅELSE AV SKADE

Sentralt i ingeniørenes samfunnsansvar står spørsmålet om potensiell skade og hvordan dette kan forebygges. Det overordnede prinsippet i mange retningslinjer handler om å unngå skade, men denne ideen må konkretiseres for å kunne omsettes i praksis. Noen dokumenter definerer unngåelse av skade fra KI-baserte systemer ved bruk av normativt språk, f.eks. at KI ikke bør militariseres eller at all KI må ha en av-knapp. Disse uttalelsene fungerer som et godt utgangspunkt. Hvordan kan vi sørge for at KI-systemer ikke gjør verden mer utrygg? Her er det kritisk å ha i mente at teknologisk utvikling virker asymmetrisk: Når teknologier har en negativ slagside påvirkes sårbare grupper langt mer og i langt større omfang. I likhet med enhver annen type teknologi kan KI forårsake fysisk, psykologisk, sosial og/eller økonomisk skade. Eksempler er beslutninger om banklån fattet av KI eller skolekretser som sier opp dyktige lærere fordi disse er blitt definert som udugelige av KI (noe som har forekommet i USA). I lys av slike eksempler stiller vi spørsmålet: Veier fordelene opp for risikoene forbundet med KI? Og hvis fordelene ikke veier opp for risikoen, bør KI overhodet utvikles? Bør vi i det minste vurdere å roe ned utviklingstempoet?

Ideen om skade er kjernen i diskusjonen om mulighetene og risikoene forbundet med KI fra et menneskerettighetsperspektiv.[34] Når man tar opp risiko og skadepotensialet kan det å betrakte fenomenet i lys av menneskerettigheter ofte gi bekymringene som fremsettes moralsk legitimitet. Likevel er det ofte vanskelig å vite hvordan man skal håndtere ideen om skadepotensial i praksis. Noen etikere har argumentert at svaret ligger i å anvende et dydsetisk perspektiv. Det vil si å oppfordre ingeniører til å dyrke etisk klokskap ved å ta hensyn til den moralske betydningen av automatiske beslutninger.[35] Men det er ikke alltid like enkelt å definere hva som utgjør en dydshandling i praksis.

På ANEs hackaton så man på sektorovergripende samarbeid mellom privat og offentlig sektor og innad i organisasjoner, hele veien fra konsernsjefen til ingeniøren. Deltakerne argumenterte for at en må få på plass helhetlige strategier på organisasjonsnivå for å unngå skade. Det er viktig at

utviklerne får muligheten til å stille seg selv «de rette spørsmålene» før og under design og utvikling av applikasjoner. Offentlige institusjoner, regelverk og standarder kan bidra i beslutningsprosesser for å unngå feiltolking av retningslinjer, og for å bidra til minimering av utilsiktet skade. Deltakerne på hackatonen var imidlertid aller mest opptatt av behovet for flere diskusjonsarenaer. Her ønsket bidragsyterne seg en mulighet for fagfolk og deres organisasjoner til å tydelig definere hva som er rett og hva som er galt innen KI-implementering. Dersom man skal etablere standarder som skal følges, bør disse utformes i fellesskap.

HÅNDTERING AV SYSTEMATISKE SKJEVHETER OG FORDOMMER

Det problemet som nok volder størst uro i forbindelse med implementering av autonome systemer er skjevheter som kan videreføre og forsterke fordommer. Algoritmer benyttes i økende grad til å styre beslutninger fattet av mennesker som er eksperter, herunder dommere, leger og ledere. Forskere og politiske beslutningstakere er imidlertid bekymret over at disse systemene uforvarende kan styrke sosiale fordommer. Det finnes dem som påstår at KI er robust nok til å kunne stå imot ekstern manipulasjon, dvs. menneskelig følelsesmessig manipulasjon, og at dette kan være en fordel, spesielt i områder som er gjennomsyret av spesielt ødeleggende menneskelige fordommer.[36] Men det er ikke lenger hold i denne argumentasjonen, særlig etter at GAN-algoritmene kom på banen. Denne utviklingen gjør at én type implementering av KI nå i praksis kan lure andre slike implementeringer ved hjelp av manipulasjonsformer som er usynlige for mennesker.[37]

Uroen rundt fordommer er forankret i det demokratiske engasjementet for å sikre rettferdige og redelige samfunn. Når skjevheter som fører til fordommer bygges inn i og reproduseres av KI-teknologier kan de ha en negativ innvirkning på særlig sårbare grupper – og slik opprettholde eksisterende sosiale forskjeller. Etter hvert som man har skjønt at slike følger kan bli en realitet, har ingeniører respondert med å utvikle et vell av konkurrerende matematiske definisjoner for hva som er en rettferdig algoritme. Nesten alle de dominerende definisjonene av rettferdighet begrenser seg imidlertid til formelle spesifiseringer som er avhengige av presise begrepsdefinisjoner. Disse igjen er noe som i all hovedsak defineres av samfunnet. Slike definisjoner reproduserer vanskelig sporbare svakheter som kan føre til alvorlige negative konsekvenser når de implementeres teknisk som objektive løsninger (Corbett-Davies & Goel).[38] Problemet med skjevheter og fordommer er at ikke alle fordommer bør elimineres. Mennesker har en rekke faste oppfatninger som er svært nyttige og som styrer menneskelig atferd hver dag. For eksempel er de fleste av oss forutinntatte mot å ta i veldig varme ting med bare hender eller å hoppe fra store høyder uten fallskjerm. Man kan hevde at slike fordommer er helt grunnleggende for vår overlevelse. Andre fordommer kan det være fornuftig for samfunnet som helhet å motvirke. To eksempler er rase- og kjønnsfordommer – det er vanskelig men helt nødvendig å imøtegå disse. Når autonome systemer utvikles, må forutsetninger og personlige fordommer granskes nøye. Det er fordi disse kan være styrende for ingeniørenes beslutninger, dog på en måte det kan være vanskelig å få tak på, noe som kan resultere i systemer som kodifiserer fordommer slik at de blir praksis.

Et tilbakevendende tema er bruken av treningsdata som inneholder skjevheter i en rekke KI-implementeringer. Igjen og igjen har man observert utfordringer med problematiske skjevheter, også i KI-implementeringer som er tenkt utrullet bredt i samfunnet. I 2016 for eksempel annonserte Microsoft at selskapet hadde utviklet en KI som ville være i stand til å bedømme fysisk skjønnhet i mennesker. Selskapet ville derfor avholde en skjønnhetskonkurranse hvor roboter skulle stå for

dømmingen. Dessverre viste resultatene en mistenkelig tendens til å likestille lys hudfarge med skjønnhet. For kort tid siden støtte Amazon på en liknende problematikk og måtte trekke tilbake et automatisert HR-system som nedgraderte alle CV-er med ordet «kvinne- (fotballag, debattlag, osv.)» i og med at slike ble definert som ikke kvalifisert for tekniske jobber. Få uker før hadde Amazon tråkket i et PR-vepsebol når ansiktsgjenkjenningssystemet som firmaet markedsførte overfor politidistrikter kom i skade for å feilklassifisere afroamerikanske senatorer som kriminelle under systemtestingen.[39]

Eksemplene er uheldige, men nokså typiske for de problemene som kan oppstå i forbindelse med KI og algoritmer. Det finnes mange grunner til at automatiserte systemer fortsetter å produsere såpass problematiske resultater. Noen av skjevhetene kan spores tilbake til treningsdataene utviklerne benyttet når de utarbeidet sine modeller. Det er vanlig at datakildene som benyttes enten hentes fra offentlige kilder eller bygger på allerede eksisterende informasjon. Problemet med å bygge modeller ut fra historiske datasett, som treningsdata, er at systemene som trenes opp med disse ender opp med å reprodusere skjevheter og fordommer – og er overraskende konsekvente når de gjør dette. Problemet er ikke nytt, men har vært et åpent spørsmål helt siden man først tok i bruk statistisk analyse til beregning av lånerisiko på 1940-tallet i USA. Den teknologiske utviklingen har aktualisert disse velkjente problemene. Disse problemstillingene er heller ikke begrenset til KI. Tvert imot, menneskelig beslutningstaking er like utsatt for mange av de samme fordommene. Ved å ta tak i disse utfordringene i sammenheng med KI kan vi bidra til å skape mer rettferdige og demokratiske samfunn.

På hackatonen uttrykte ingeniørene bekymring ved tanken at fordommer er en iboende om enn utilsiktet egenskap i automatiserte systemer. Inndata og (cluster) algoritmer ble identifisert som de viktigste årsakene til fordommene. Ettersom det til syvende og sist er vanskelig eller umulig å eliminere fordommer fullstendig, setter løsningene som ble foreslått og som er gjengitt nedenfor søkelyset på bevisstgjøring og granskning av KI-løsninger.

Avdekking av fordommer
I situasjoner hvor skjevheter ikke er ønskelige er det hensiktsmessig å sørge for at eventuelle skjevheter avsløres og – ved behov – håndteres. Det er viktig å anerkjenne det snevre fokuset til teknisk utvikling og dens resultater. Som en av deltakerne på hackatonen uttalte: «Tech-gutta (vi ingeniører) i bransjen vet ikke alltid hvor og hvordan data samles inn og forbehandles. Også når vi nærmer oss sluttresultatet, vet vi ikke alltid hvor resultatet skal benyttes. Dette er fordi det finnes salgs-/markedsføringsavdelinger og KI som settes i produksjon men hvor systemets fulle omfang ikke er kjent.» Hvordan kan ingeniører avdekke potensielle fordommer i datasettene de benytter i de tilfeller hvor dataenes opphav er vanskelig å fastslå, og der hele omfanget av systemet ikke er synlig for dem? Jo større og mer kompleks organisasjonen er, jo mer aktuelle kan vi anta at slike problemer blir.

Granskning av systematiske skjevheter og fordommer
Mange deltakere kommenterte at det bør finnes en mulighet for å få gransket KI-systemer gjentatte ganger av eksterne, nøytrale enheter (f.eks. et «modelltestingsinstitutt»). Slike systemer burde så testes på nøye utvalgte data og kun få godkjenning dersom ingen større skjevheter avdekkes. En ingeniør kom med følgende kommentar: «Man kunne ha innført et krav om godkjenning fra en uavhengig enhet for KI-systemer som benyttes i forvaltning, offentlige helsetjenester og på utdanningsinstitusjoner. Dessuten burde det gis innsyn i kriteriene for modelldesignen, selve modellene (med mindre dette hindres av personvernhensyn) og resultatene av evalueringen av

modellen.» Mange deltakere skilte mellom systemer som er sentrale deler av samfunnets maskineri og systemer som kun er innrettet mot reklame og produksjon og kjøp av forbruksvarer. Førstnevnte gruppe systemer bør det føres mer tilsyn med enn sistnevnte, ble det slått fast. Oppgaven med å avdekke fordommer er for viktig til å overlates til den enkelte ingeniøren. Det var bl.a. fordi avdekking av skjevheter og beslutninger om hvilke skjevheter som ikke bør få plass i forvaltningen, helsetjenester eller utdanningssystemer faller utenfor domenet av det å være rene tekniske avgjørelser. Dette er beslutninger som bør fattes av styresmaktene.

Det ble stilt spørsmål om det er akseptabelt å produsere systemer der fordommene er kjent og hva man kan gjøre for å redusere slike utfordringer. Her er det helt klart et spørsmål om å utøve etisk og moralsk skjønn, og enkeltpersoner – helt uavhengig av hvor mye integritet hver enkelt måtte ha – må få støtte av sine omgivelser for å utvikle et standpunkt. Hvilke skjevheter er akseptable og hvilke skjevheter ønsker de nordiske landene å unngå? Det er hvert lands prioriteringer og moralske idealer som avgjør hvor mye man skal satse på dette arbeidet.

Arbeid med de viktige spørsmålene

IEEEs uttalelse om «Ethically Aligned Design» gir uttrykk for generelle prinsipper vedrørende utviklingen av alle typer KI – autonome og intelligente systemer (K/IS), uten hensyn til om det er snakk om fysiske roboter eller programvare.

AUTONOME/INTELLIGENTE SYSTEMER MÅ

1. *Romme de høyeste idealene for menneskelig godhet som overordnede menneskerettighetsverdier*.

2. *Gi prioritet til at bruken gagner menneskeheten og naturen. Merk at det ikke er noe motsetningsforhold mellom disse to. Menneskeheten og naturen står i et gjensidig avhengighetsforhold. Prioritering av menneskelig velferd skal ikke føre til en forringelse av miljøet.*

3. *Redusere risiko og negative virkninger, herunder misbruk ettersom autonome/intelligente systemer utvikles som sosiotekniske systemer. Et viktig element i så måte er å sørge for at autonome/intelligente systemer er transparente, og at ansvarsforholdene er klare.*

Disse generelle prinsippene underbygges av ingeniørers og andre interessenters ansvar som er omtalt i en rekke andre, tilsvarende dokumenter. Disse inkluderer kravet om å vurdere prioriteringene og å sikre at menneskets interesser går foran kjerneinteressene til institusjoner og kommersielle aktører. I tråd med prinsippene for menneskesentrert databehandling har mange nyere yrkesetiske retningslinjer fokus på å sette mennesker i sentrum for teknologidesign, og vektlegger menneskesentrert design og ingeniørpraksis. Retningslinjene til Association of Computing Machinery er et eksempel på dette. Det påpekes fra flere hold at en slik vektlegging er en forutsetning for at allmennheten skal få tillit til KI-systemer. Tillit kan opparbeides over tid og gjennom naturlige interaksjonsmodaliteter, men står også i fare for å lett undergraves gjennom skjødesløs databehandling eller uforståelige beslutninger som påvirker folks liv. Behovet for å systematisere beste praksis er en gjenganger i litteraturen. Dette er en forutsetning for å utvikle KI-teknologier og opprettholde tilliten til dem. En slik samling retningslinjer for beste praksis kan fungere som en veileder for en trygg og etisk utvikling og håndtering av KI. Denne samlingen må være en nøye gjennomtenkt sammenstilling av sosiale normer og verdier, algoritmisk ansvarsstillelse, etterlevelse av eksisterende lovgivning og politikk, og krav til personvern og persondatavern.

På grunn av KI-systemers evne til å samle og raskt behandle enorme datamengder, nevnes personvern som en av de største bekymringene. Mange KI-teknologier muliggjør innsamling, oppfølging og utveksling av personopplysninger raskt, rimelig og i mange tilfeller uten at menneskene som berøres av dette får vite om det. Man er derfor opptatt av å sørge for at fagfolk innen IT blir gjort godt kjent med de ulike definisjonene og former for personvern. De bør forstå hvilke rettigheter og ansvar som er forbundet med innsamling og bruk av personopplysninger. EUs generelle personvernforordning (EU GDPR) krever at man anvender innebygd personvern (på engelsk «data protection by design» som forkortes til DPbD) når man utvikler dataintensive systemer.

Med bakgrunn i den omfattende debatten rundt skaden som KI-systemer kan forårsake, er det naturlig å møte denne utfordringen. Det er mange som argumenterer for at man skal være ekstra

nøye med å identifisere og minimere potensielle risikoer i maskinlæringssystemer. Systemer hvis fremtidige risiko ikke kan forutsies på en pålitelig måte krever hyppige risikovurderinger ettersom

systemet utvikler seg ved bruk – ellers bør det ikke tas i bruk. Eventuelle sider ved systemet som kan medføre stor risiko for skade må meldes inn til rett aktør. KI-systemer bør innbefatte forklaringsbaserte parallelsystemer eller muligheten til tilbake rulling av beslutninger, slik at direkte konsekvenser kan reverseres.

En av bekymringene med autonome og selvlærende algoritmer er hvordan de brukes i utviklingen av autonome våpen. Den ideelle organisasjonen Article 36 har utarbeidet prinsippet «meningsfull menneskelig kontroll» for å uttrykke selve kjernen i det som trues av utviklingen i retning av økt autonomi i våpensystemer. Dette prinsippet krever at menneskelig skjønn må utøves på en måte som har faktisk betydning når autonome våpen og andre kritiske systemer skal brukes.

Utviklingen av dataintensive KI-systemer skaper naturlig nok et bredt spenn av potensiell skade. Et eksempel er personvernutfordringene som følger av hvor raskt og bredt data kan samles inn ved hjelp av KI-systemer. Dette er spesielt urovekkende i situasjoner hvor menneskerettigheter står sentralt. Man bør derfor være særlig oppmerksom på sårbare mennesker, f.eks. slike personer som av politiske, økonomiske, sosiale eller helsemessige grunner er særlig utsatte for profilering som kan redusere deres selvbestemmelse eller som kan utsette dem for diskriminering eller stigmatisering. Å hensynta sårbare elementer innebærer også å arbeide aktivt for å redusere fordommer i utviklingen av selvlærende algoritmer.

Ved å definere rettslige grenser for klassifisering og beslutninger kan de som blir berørt gjøres kjent med at de har å gjøre med en intelligent maskin. Dette er helt sentralt, spesielt i møte med mindre privilegerte og sårbare grupper. Ettersom implementering av KI – på grunn av disse teknologiers bruk av historiske data – innebærer en risiko for at gamle og ofte utdaterte oppfatninger befester seg, er det viktigere enn noen gang å ta tak i spørsmål som rettferd og ansvarsstillelse. Dette krever utforming av nye modeller for rettferdig fordeling og godefordeling etter hvert som automatisering, digitalisering og KI fører til økonomiske endringsprosesser. Det krever også at man sikrer at det gis innsyn i sentral KI-teknologi og tilrettelegger for opplæring innen realfag og digitale fag. Videre innebærer dette at man må utvise økt årvåkenhet overfor prosesser som kan undergrave sosial samhørighet, fostre radikal individualisme, true, hemme eller påvirke politisk beslutningstaking, innskrenke ytringsfriheten og retten til å gi og få opplysninger uten innblanding.

Det er opplagt at disse hensynene er av stor betydning. Men hva skal ingeniører gjøre dersom de blir oppmerksomme på noen av disse problemene? Det bør tilrettelegges for rapporteringsmuligheter dersom man støter på tegn på systemrisiko som kan forårsake skade. Ledere bør prioritere begrensning av de risikoene som avdekkes og iverksette tiltak for å redusere eventuelle skader. I de tilfeller der man unnlater å iverksette slike tiltak kan det bli nødvendig å «varsle», med sikte på å unngå potensielle skadevirkninger. For å bidra til dette bør systemene bygges slik at det gis mulighet til å innrapportere tilbakemeldinger, relevante forklaringer og inngi klager.

Disse og andre vurderinger om hvordan man skal håndtere kommende problemer er blitt mye diskutert i mange aktuelle dokumenter, og gjennom hele ANEs hackaton, hvor deltakerne diskuterte behovet og gjennomførbarheten av mange av løsningene som ble foreslått. Disse debattene utgjorde grunnlaget for en første versjon med retningslinjer for ingeniørarbeidet samt anbefalinger for tiltak fra myndighetens side.

ETIKK I INGENIØRUTDANNINGEN?

Hva innebærer det å være en ansvarlig ingeniør? Hvordan kan ingeniører finne ut hva som er ansvarlig og hva som er uansvarlig atferd, og hva er egentlig ingeniørenes ansvar? Det er opplagt at det første møtet med slike spørsmål bør finne sted under ingeniørutdanningen. Praksis som studenter lærer og tilegner seg under studiene vil videreutvikles gjennom hele yrkeslivet.

Som Google hevder: ingeniører må «dele kunnskap om KI på en ansvarlig måte ved å publisere opplæringsmateriell, beste praksis og forskning som gir flere mennesker muligheten til å utvikle nyttige KI-applikasjoner».[40] I tillegg til ønsket om å gi flere mennesker muligheten til å utvikle KI-applikasjoner, er det blitt uttrykt bekymring fra mange hold for hvordan man kan gi folk muligheten til å forstå eksisterende KI-applikasjoner uten at de har omfattende kunnskap om informatikk eller datametoder. Hva må endres innen ingeniøropplæringen på de ulike nivåene? Hva mangler, og hva må det tas tak i? Praktiserende ingeniører er de som er best egnet til å begynne å svare på disse spørsmålene.

Mange ingeniører som arbeider med KI i gründerbedrifter eller i etablerte bedrifter sliter fordi – som en ingeniør forklarte det under ANE-hackatonen: «Foreløpig vet vi ikke hva som forventes av de som designer, utvikler og benytter KI-systemene. Ansvarsforholdene og ansvar er ikke klart definert.» Det finnes med andre ord foreløpig ingen reell presedens eller erfaringer som kan vise hva som er «god, ansvarlig atferd». Selv om ingeniører prøver å opptre ansvarlig, er spørsmålet om hva som utgjør ansvarlighet i praksis fortsatt et komplisert spørsmål med mange ukjente variabler. En gruppe ingeniører lurte for eksempel på om det overhodet er mulig å ha fordomsfrie datasett for maskinlæring samt hvordan man skal identifisere eventuelle skjevheter. De var klare over at en måtte være kritisk mht. treningsdata – men det hersket usikkerhet rundt hva de skulle se etter og hva som utgjør skadelige skjevheter og fordommer.

Et løsningsforslag som mange ingeniører støttet var å samle ingeniører og relevante interessenter for å utvikle ideer om hva ansvar betyr i denne sammenhengen. Viktige spørsmål er bl.a.: Hvilke forhold må tas med i betraktning, hvilke forpliktelser bør man påta seg og bør lovfestes, nettopp fordi bygging av nye systemer krever at man anerkjenner og reforhandler innbyrdes ansvarsforhold. Samtidig påvirkes de beslutningene som fattes i stor grad av endringer i normer og nye regler. Det er også relevant å spørre seg hvem som får ta disse beslutningene og av hvilke verdier de styres. I en globalisert økonomi fungerer ikke forestillingen om «godt» som et lokalt begrep – samtidig er «godt» alltid kontekstavhengig. Hvem har da ansvaret når det som er «godt» tar en helomdreining og får negative følger?[41] Slike diskusjoner får mange ingeniører til å reflektere over hvordan ingeniørutdanningene bør endres helt fra startgropen, slik at fremtidige ingeniører kan utveksle erfaringer om dette enda tidligere. Håpet er at unge ingeniører kan utvikle en modnere etisk tilnærming i sitt virke.

Spørsmål om KI og utdanning dreier seg om mer enn ingeniørutdanningen. Det finske departementet for økonomi og arbeidsliv har uttalt: «Det er et behov for større allmennkunnskap innen KI. Folk trenger å forstå hvordan ting kommer til å fungere i KI-alderen». Spørsmålet her er hva slags allmenndannelse samfunnet trenger som helhet. Hvilke grunnleggende begreper bør alle kjenne til – og er det mulig å oppnå dette?

INSTITUSJONENES ANSVAR

Hvem skal påta seg ansvaret som er blitt nevnt i debattene? De nordiske landene utmerker seg med utstrakt tillitt til staten og høye forventninger til at næringslivet skal opptre etisk. Utover det som er enkeltpersoners ansvar, hvem skal påta seg de nye ansvarsområdene, og hvordan skal disse defineres? Hvilken rolle bør bransjeorganisasjoner som ANE eller nasjonale fagforbund ha med

hensyn til å støtte ingeniører som søker å opptre ansvarlig? Hvilke forpliktelser har ingeniørenes arbeidssteder? På hvilke måter bør disse organisasjonene endre seg? Hvilke forpliktelser har styresmaktene mht. etikk og KI?

Anbefalinger

Selv om ingeniører og deres organisasjoner vil måtte håndtere mye av ansvaret innen design og implementering av KI-systemer, må de aktuelle styringsorganene i de nordiske landene og på EU-nivå ta inn over seg sitt eget ansvar og handlingsrom. I de tilfellene der ingeniørenes etiske praksis bør overlates til selskapene og ingeniørene selv, bør spørsmål som nødvendige endringer i utdanning og implementering av nye typer lover og regler også i framtida må være et nasjonalt og regionalt ansvar. I denne forbindelsen, legges det herved frem et sett med anbefalinger for vurdering.

STRATEGISKE ANBEFALINGER

1. *Diskusjonen må forankres på politisk nivå. Videre må offentligheten få en bedre forståelse av KI. Et bidrag i den forbindelse kunne vært å etablere en plattform, nærmere bestemt en arena der beslutningstakere, næringslivet, forskningsmiljøene, sivilsamfunnet og fagfolk, herunder ingeniører, gis muligheten til å møtes for å utvikle robuste og transparente KI-løsninger gjennom felles samtaler.*

Det er klart at det å imøtegå potensielle utfordringer i forbindelse med bred implementering av KI-teknologier krever statlige tiltak og myndighetstilsyn. Samtidig omfatter de konkrete problemene som oppstår i forbindelse med autonome systemer også teknologiske sider som krever høy teknisk spisskompetanse for å kunne utvikle løsninger og lovforslag som både kan bidra til å fremme innovasjon samtidig som man imøtegår potensielle utfordringer. Spørsmålet her er hvordan vi kan ta ut nytteverdien i KI-teknologier uten å utnytte brukerne. Styresett basert på offentlig engasjement står sentralt i nordisk kultur. Denne tilnærmingen åpner for å involvere ulike former for spisskompetanse i styresmaktenes evalueringsprosesser. Utviklingen av nye former for engasjement av denne typen forutsetter imidlertid politisk vilje og at det satses økonomisk.

2. *Både i tekniske fag og i arbeidslivet er utdanningstilbudet innen etisk skjønn og etiske retningslinjer ofte mangelfullt. Dette må det tas tak i ved å endre opplæringsmålene og prioriteringene innen tekniske fag. Videre må det tilrettelegges for relevante muligheter for livslang læring.*

Samtaler om etiske spørsmål forbundet med implementeringen av KI-systemer forutsetter et godt utviklet begrepsapparat og ikke minst kjennskap til eksisterende etiske rammer og deres begrensninger. Forsøk på å utvide eller til og med endre opplæringen innen tekniske fag er alt igangsatt på flere nivåer – helt fra å introdusere grunnleggende etikkmoduler i tekniske kurs til utvikling av nye workshoper og kurs. Mye av denne utviklingen har sin bakgrunn i innsats på grasrotnivå og forankring i sivilsamfunnet og kommersielle aktører. For at disse endringene skal bli systemomfattende er det viktig med støtte og tilsynsvirksomhet fra myndighetenes side.

3. *Det er viktig å etablere en offentlig klageordning. En slik prosess må sette enkeltpersoner og organisasjoner i stand til å ta opp KI-atferd og -beslutninger som de vurderer som potensielt skadelige.*

En av de største bekymringene mht. KI-teknologier er spørsmålet hvordan folk som berøres av eventuelle feil (noe som allerede har skjedd i mange tilfeller) skal bli gitt muligheten til å reagere på en måte som ivaretar deres aktørskap og verdighet. Ansvarlige organisasjoner må satse på å etablere tydelige ansvarsledd og klare ansvarsforhold i hele levetiden til de aktuelle tekniske systemene samt støtte dialogen med berørte grupper. Slike prosesser forutsetter godkjennelse og støtte fra

myndighetene, samt godt organiserte tilsynsformer for å sikre at konsekvensene blir tydelige og for å bygge tillitt.

4. *Det er nødvendig å utarbeide et regel- og lovverk for KI-relaterte problemstillinger som formelt definerer og regulerer ansvarsforhold.*

Det er klart at de som designer og utvikler teknologier må holdes ansvarlige for sine beslutninger og handlinger. Men dette kan kun oppnås dersom vi anerkjenner at både den enkelte ingeniør og de organisasjonene den enkelte ingeniører hører til er en del av de sosiale, politiske og økonomiske systemene som utgjør samfunnet vårt. Det er viktig å formalisere ansvarsforhold og definere hvem som skal stilles til ansvar når ting gjør ting og dette fører til negative konsekvenser.

5. *Ingeniører, politiske beslutningstakere, sivilsamfunnet og offentligheten trenger arenaer der den aktive dialogen rundt KI-relaterte etiske problemstillinger kan utspille seg. Det må legges til rette for utviklingen av slike arenaer gjennom finansiering og annen støtte.*

Det er et akutt behov for refleksjon rundt hva som er etikk mht. KI og hvordan man kan avgjøre hva som er rett og galt når KI implementeres og får følger. Slike rom for refleksjon bør gi fagfolk og beslutningstakere med ulik bakgrunn og ulike spesialfelt muligheten til å møtes og diskutere. Det er ikke kun opp til relevante interessenter å støtte slike refleksjoner og diskusjoner. For å lykkes, kreves det fortløpende politisk støtte og innsats fra myndighetenes side.

ANBEFALINGER OG RETNINGSLINJER FOR INGENIØRER OG DERES INSTITUSJONER

Selv om intensjonen med dette dokumentet er å tale direkte til ingeniørene, må vi anerkjenne følgende to behov:

1. *Den enkelte ingeniør må ha nødvendig utdanning og opplæring for å kunne ta sitt ansvar.*
2. *For at ingeniører skal kunne påta seg sitt ansvar og de problemstillingene som følger av kunstig intelligens på best mulig måte, må ingeniører ha støtte fra de organisasjonene og institusjonene som de samarbeider med og arbeider for.*

Anbefalingene er utviklet ut fra diskusjoner med ingeniører og på grunnlag av en oversikt over andre aktuelle satsinger for å takle KI-relaterte etiske problemstillinger. Det er ikke bare opp til den enkelte ingeniøren å følge disse retningslinjene. Dette er fordi etisk utvikling av KI krever mer enn at enkeltpersoner påtar seg ulike typer etisk ansvar. Det foreligger allerede et vell av retningslinjer for ingeniører som beskriver etisk atferd. Flere av retningslinjene under kan følges både av enkeltpersoner og av organisasjoner sammen med de retningslinjene som allerede finnes i landene i Norden. Mange av disse retningslinjene retter seg imidlertid mot organisasjonspraksis og ikke enkeltpersoner. Dette er fordi arbeidet for å styrke etisk praksis krever solid forankring i institusjonene for å være effektivt. Derfor kreves det at organisasjonene engasjerer seg i arbeidet med å takle etiske problemstillinger i forhold til KI.

Anbefalingene legges frem ut fra en forståelse av at vellykket implementering krever både innsats og engasjement fra den enkelte ingeniøren i tillegg til alle deres organisasjoner i fellesskap.

ANBEFALINGER FOR Å FREMME ETISKE VURDERINGER VED UTVIKLING OG IMPLEMENTERING AV KUNSTIG INTELLIGENS

1. *Skape arenaer for debatt rundt KI-relaterte problemstillinger og etikk. Både arbeidsplasser og organisasjoner i sivilsamfunnet må tilrettelegge for å støtte slike arenaer.*

Refleksjonsprosessene med å definere KI og etikk viser at det er nødvendig å støtte diskusjoner som tar opp KI og etikk. Slike utvekslinger må gis tilstrekkelig rom og tid slik at de kan modnes før de kan anvendes i praksis. Selv om ANEs hackaton var vært innrettet slik at man har konsentrert seg om ANEs medlemmer, er det viktig å få til slike debatter også andre steder og med andre deltakere – blant eksperter, interessenter og på langt bredere plan, med samfunnsmedlemmer.

2. *Investere i og utvikle verktøy som tilrettelegger for etisk debatt, spørsmål og beslutningstaking gjennom hele designprosessen – og ikke bare i prosessens start- og sluttfase.*

Det å ta etiske avgjørelser er ikke noe som skjer kun på begynnelsen og slutten av en prosess, og det er mer enn et ekstrakrav man må innfri. Dette er fordi etiske spørsmål kan føre til at man stiller spørsmål ved selve grunnlaget for det ferdige arbeidet eller gjenstanden som er produsert – noe som kan gjøre at hele prosjektet kan vise seg å være uforenelig med etiske rammeverk. Etisk debatt og vurdering er med andre ord ikke noe som kan legges inn i prosjektets startfase og sluttfase som et «kontrollpunkt» for å sikre at det man har produsert kan beskrives som «etisk». Det er heller slik at en må integrere den etiske tilnærmingen gjennom hele design- og utviklingsprosessen som en metode som kan styre prosjektet fra A til Å, og ikke som et sett med leveranser.

3. *Etablere et sett med interne standarder og sjekklister som tar for seg etiske problemstillinger ved utvikling av kunstig intelligens, for å sikre meningsfull menneskelig styring.*

Det er viktig at etiske vurderinger legges frem gjennom hele prosessen, inkludert design og utvikling, men i praksis er det vanskelig å alltid gi etiske vurderinger prioritet i situasjoner der forholdene rundt prosjektet kan endre seg som følge av eksterne faktorer. Ved å utforme et sett med interne normer og sjekklister for å takle etiske spørsmål kan en redusere utfordringen med å alltid opprettholde det etiske engasjementet. Et slikt sett kan fungere som et brukervennlig verktøy for å uttrykke spørsmål på en måte som er relevant for prosjektet eller oppgaven det gjelder. Ved å eksempelvis inkludere et punkt på sjekklisten som påpeker at det er viktig å sikre meningsfull menneskelig kontroll kan prosjektdeltakerne vende tilbake til dette spørsmålet flere ganger i lys av nye aspekter etter hvert som de utvikles i prosjektet. Selv om meningsfull menneskelig kontroll som begrep er blitt tatt i bruk for å vurdere autonome våpen, benytter vi det her som et langt bredere konsept som skal sikre et visst nivå menneskelig kontroll i samhandling med alle KI-systemer.

4. *Støtte og tilrettelegge for intern rapportering av risiko og eventuelle brudd på retningslinjer, og etablere rutiner for tiltak og oppfølging.*

Etiske retningslinjer i seg selv kan bli en formalitet, en boks som må krysses av, uten at dette innebærer noen endring i selve prosjektet. For å hindre denne typen utfall og for å gi prosjektdeltakere mulighet til å uttrykke bekymringer de opplever, må det etableres kanaler for å

rapportere risiko og retningslinjebrudd innenfor de aktuelle institusjonene, samt fastsettes tydelige tiltak og reaksjoner dersom brudd finner sted.

5. *Opprette interne opplæringsprogrammer for medarbeidere med sikte på å skape en dypere forståelse av etikk, og å utvikle evnen til etisk refleksjon, debatt og til å identifisere systematiske skjevheter.*

Interne opplæringsprogrammer kan gi deltakere muligheten til å teste sine ideer og gi dem muligheten til å skape sine egne former for etisk engasjement. For eksempel, dersom fordommer viser seg å være uunngåelige, kan disse håndteres gjennom opplæring som gir mennesker muligheten til å kjenne igjen og imøtegå dem. Slike tiltak er også et synlig bevis på at institusjonen er villig til å la sine arbeidstakere bruke tid og ressurser på utvikling mht. etisk refleksjon. Etiske og moralske resonnement forutsetter opplæring og anvendelige rammer.

6. *Ha særlig fokus på mulige systematiske skjevheter som kan oppstå under utvikling av systemer, i treningsdata og modellenes funksjonalitet. Spesielt de som kan påvirke særlig utsatte personer.*

Gitt hvor mye oppmerksomhet som for tiden gis treningsdata i utvikling av alle KI-systemer som benytter algoritmisk databehandling, er det viktig å sørge for at disse hensynene imøtegås i praksis. Ved å lære å tenke ikke bare ut fra gjennomsnittsverdier men også ekstremtilfeller kan man bidra til oppvurdering av effekten på de mest sårbare gruppene – ut fra designspesifikasjoner. Dette fører igjen til kreative beslutninger og bedre løsninger.

7. *Utvikle aksept for virksomhetsansvar for potensiell skad. For eksempel ved å etablere kanaler for å håndtere skade påført andre av KI-systemer som virksomheten har designet.*

Hvordan kan man gi mennesker som berøres av KI-systemer muligheten til å reagere dersom ting ikke går som planlagt? Dersom noe går galt når et KI-system implementeres og folk blir skadelidende (uavhengig av om de samhandlet med systemet direkte), er det et spørsmål om hvem som skal ta ansvaret for de negative konsekvensene. Per i dag er det ikke klart hvordan mennesker som påvirkes negativt bør reagere, hvem de burde kontakte og hvem som skal bistå dem ved spørsmål. Slik usikkerhet avler mistillit og sår tvil rundt nytteverdien av KI-systemer, noe som fører til avvisning fremfor aksept. For å opprettholde tillit er det viktig å etablere tydelige ansvarsledd og ansvarsforhold gjennom hele levetiden til teknologisk systemer.

8. *Etablere en intern etisk evalueringsprosess som bidrar til å gjøre selskapers beslutningsprosesser mer demokratiske ved å involvere flere interne aktører.*

Interne etiske granskninger gir ikke bare mer stabil forankring for evaluering av løpende satsinger innenfor organisasjoner, men gir også bekymrede ansatte – som ikke alltid har muligheten til å uttrykke sine bekymringer – sjansen til å delta i beslutningstaking. Slike evalueringsprosesser kan også tilrettelegge for at ideer og erfaringer deles på tvers av hele organisasjonen.

9. *Arbeide for økt transparens, ikke bare i beslutningene som fører til design og utvikling av KI-systemer, men også i virksomhetenes ansvarskjede.*

Ved å synliggjøre egne ansvarsledd kan organisasjoner vise at de forplikter seg til å etablere mekanismer for ansvarsstillelse ved eventuelle skadehendelser. Dette betyr ikke at økt transparens i beslutningstaking, design eller utvikling ikke er ønskelig. Tvert imot betyr dette at disse to prosessene utfyller hverandre – dersom en av dem mangler kan dette skade tilliten til organisasjonen.

10. *I arbeidet for transparens, opprettholde en bevissthet om at transparens også kan by på etiske fallgruver og begrensninger.*

Selv om transparens er en god målsetting for organisasjoner som designer og utvikler KI-teknologi, må det etiske engasjementet uttrykkes på flere måter. Transparens er kun én del av likningen og denne rapporten har også reflektert rundt mange andre bekymringer som det må tas høyde for. En annen side ved transparens er at dersom transparens ikke følges av ansvarsstillelsesmekanismer – f.eks. når algoritmer benyttes for å ta diskriminerende beslutninger – kan det bli svært vanskelig å oppnå virkelig endring.

Kilder

[1] Agre, «Toward a Critical Technical Practice: Lesson Learned in Trying to Reform Al.»

[2] Khalil, «Artificial Decision-Making and Artificial Ethics.»

[3] Crawford og Calo, «There Is a Blind Spot in Al Research.»

[4] Irani, «The Hidden Faces of Automation.»

[5] Agrawal, Gans, og Goldfarb, «The Simple Economics of Machine Intelligence.»

[6] Bird m.fl., «Exploring or Exploiting?»

[7] Yampolskiy, «Artificial Intelligence Safety Engineering.»

[8] Silverstone, «Proper Distance: Toward an Ethics for Cyberspace.»

[9] Bostrom og Yudkowsky, «The Ethics of Artificial Intelligence.»

[10] European Group on Ethics in Science and New Technologies, «Artificial Intelligence, Robotics and 'Autonomous' Systems.»

[11] European Group on Ethics in Science and New Technologies

[12] IEEE Standards Association, «Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems.»

[13] Pichai, «Al at Google: Our Principles.»

[14] Rossi, «Artificial Intelligence: Potential Benefits and Ethical Considerations.»

[15] Burrell, «How the Machine Thinks.»

[16] Wagner, «Ethics as an Escape from Regulation.»

[17] Brey, «Anticipating Ethical Issues in Emerging IT.»

[18] Howard og Borenstein, «The Ugly Truth About Ourselves and Our Robot Creations.»

[19] NITO, «Etikk for ingeniører og teknologer.»

[20] BS 8611: 2016, «Robots and Robotic Devices: Guide to the Ethical Design and Application of Robots and Robotic Systems.»

[21] IEEE Standards Association, «Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems.»

[22] Vallor m.fl., «An Introduction to Software Engineering Ethics.»

[23] Latonero, «Governing Artificial Design.»

[24] Cath m.fl., «Artificial Intelligence and the 'Good Society'.»

[25] Cath m.fl.

[26] Cath m.fl.

[27] «Background. The Iron Ring.»

[28] IEEE Standards Association, «Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems.»

[29] Ananny og Crawford, «Seeing without Knowing.»

[30] Albu og Flyverbom, «Organizational Transparency.»

[31] Ananny og Crawford, «Seeing without Knowing.»

[32] Cath m.fl., «"Artificial Intelligence and the 'Good Society': The US, EU and UK Approach:»

[33] Simon, «Distributed Epistemic Responsibility in a Hyperconnected Era.»

[34] Latonero, «Governing Artificial Intelligence.»

[35] Shilton, «Values and Ethics in Human-Computer Interaction»; Vallor m.fl., «An Introduction to Software Engineering Ethics.»

[36] Bostrom og Yudkowsky, «The Ethics of Artificial Intelligence.»

[37] Fawzi, Fawzi og Frossard, «Analysis of Classifiers' Robustness to Adversarial Perturbations.»

[38] Corbett-Davies og Goel, «Defining and Designing Fair Algorithms.»

[39] Levin, «A Beauty Contest Was Judged by AI and the Robots Didn't like Dark Skin»; Lee, «Amazon Scrapped 'sexist AI' Tool»; Singer, «Amazon's Facial Recognition Wrongly Identifies 28 Lawmakers, A.C.L.U. Says.»

[40] Pichai, «AI at Google: Våre prinsipper.»

[41] Shiklovksi, «Responsibilty in IoT: What Does it Mean to 'Do Good'? »

Informasjon om arrangementet

Hackatonen «Nordiske ingeniørers holdning til kunstig intelligens og etikk» ble organisert av «The Association of Nordic Engineers»(ANE) og «The IT University of Copenhagen, EthosLab» i fellesskap. Arrangementet fant sted 25. september 2018, og samlet medlemmer fra ANEs medlemsorganisasjoner. Det vil si ingeniører og studenter innen informasjonsteknologi og naturvitenskap med kompetanse om kunstig intelligens(KI) og anvendelse av det.

ANE er et samarbeid mellom Sveriges Ingeniører, Ingeniørforeningen i Danmark(IDA), Verkfrædingafélag Íslands(VFI) og NITO – Norges Ingeniør- og Teknologorganisasjon. I tillegg pågår det forhandlinger om medlemskap for de finske ingeniørorganisasjonene, og de finske organisasjonene sendte en deltager til hackatonen.

Til sammen representerer ANE mer enn 340 000 ingeniører i Norden. ANEs formål er å jobbe for nordiske ingeniørers interesser gjennom nordisk samarbeid på tvers av de nasjonale organisasjonene.

Arrangementet var fasilitert av stipendiater og forskere ved «The IT University of Copenhagen, EthosLab» som spesialiserer seg i forskning på etikk og teknologisk utvikling.

MER INFORMASJON

Kortversjon av rapport på norsk.

Les mer om arrangementet på nito.no.

KONTAKTINFORMASJON

Inese Podgaiska, Generalsekretær ANE

Telefon: +4529743960

E-post: ipo@ida.dk