



SUPERINTELLIGENS, MAKT OG MENNESKERETTIGHETER

ISBN 978-82-8400-040-4 (trykket utgave)
ISBN 978-82-8400-041-1 (elektronisk utgave)

Utgitt: Oslo, desember 2025
Forside: Birgitte Blandhoel
Elektronisk publisert på: <https://teknologiradet.no>



«På vår planet er det ikke plass for to tendenser som begge forsøker å beherske verden. Den ene må vike. Vi vet alle hvilken.»

Karel Čapek: «Salamanderkrigen» (1936)

FORORD

Ideen om maskiner som kan overgå menneskelig intelligens er ikke ny, men den raske teknologiutviklingen innen maskinlæring og generativ kunstig intelligens har gitt nytt liv og aktualitet til konseptet.

Verdens ledende teknologiselskaper har ambisjoner om å utvikle superintelligens – kraftige og avanserte KI-systemer som overgår intelligensen til de smarteste menneskene i verden. Dersom de lykkes, kan det få store konsekvenser for mennesker og samfunn, og forsterke geopolitiske spenninger.

For fjerde år på rad har Teknologirådet og Norges institusjon for menneskerettigheter (NIM) kombinert vår ekspertise og perspektiver på menneskerettigheter og teknologi i en felles rapport. I år utforsker vi ulike framtidsscenarioer for et 2030 med superintelligens, og vurderer konsekvenser for menneskerettighetene avhengig av hvem som kontrollerer teknologien: selskaper, stater, globalt samarbeid eller et økosystem av ulike aktører.

Vår analyse viser at det å sikre ansvarlig utvikling og bruk av kraftige KI-systemer er en presserende demokratisk utfordring, som krever solid innsikt i teknologiutviklingen, politisk handlekraft og et sterkt vern om grunnleggende rettigheter.

Takk til Hanne Sofie Lindahl og Joakim Valevatn fra Teknologirådet, og Cecilie Hellestveit, Vidar Strømme og Vetle Seierstad fra NIM som har ledet prosjektet.

Desember 2025

Kai Spurkland
direktør
Norges institusjon for menneskerettigheter

Tore Tennøe
direktør
Teknologirådet

INNHold

VISJONEN OM SUPERINTELLIGENS 6

HVORDAN FORSTÅ SUPERINTELLIGENS?..... 6

HVOR SANNSYNLIG ER SUPERINTELLIGENS?..... 12

MENNESKERETTIGHETENE OG SUPERINTELLIGENS 19

HVEM KONTROLLERER SUPERINTELLIGENS? 20

HVILKE MENNESKERETTIGHETER KOMMER I SPILL? 22

FIRE FRAMTIDER MED SUPERINTELLIGENS 27

FRAMTIDSSCENARIOER FOR SUPERINTELLIGENS..... 27

FRAMTIDSSCENARIO 1: SELSKAPER HAR KONTROLL 29

FRAMTIDSSCENARIO 2: STATER HAR KONTROLL 34

FRAMTIDSSCENARIO 3: GLOBAL KONTROLL 39

FRAMTIDSSCENARIO 4: INGEN HAR KONTROLL 43

HVA KAN VI LÆRE AV FRAMTIDSSCENARIOENE? 46

HOVEDKONKLUSJONER 47

REFERANSER 48

VISJONEN OM SUPERINTELLIGENS

Utviklingen av kunstig intelligens går fort, og har gitt grobunn til en diskusjon om vi er på vei mot superintelligens – maskinintelligens som overgår mennesker. Dette vil kunne få stor påvirkning på mennesker og samfunn.

HVORDAN FORSTÅ SUPERINTELLIGENS?

Teknologiutviklingen går fort, og kunstig intelligens bygget på store språkmodeller blir stadig kraftigere og mer anvendelig. Spørsmålet om maskiner kan utvikle intelligens som overgår menneskelig kognitiv kapasitet og dermed utgjøre en slags superintelligens, har fram til nylig vært en filosofisk og hypotetisk diskusjon.

Teknologiens sprengkraft og raske utvikling, drevet fram av gjennombruddet for store språkmodeller, har imidlertid gitt grobunn for en langt mer konkret diskusjon om hva superintelligens vil innebære for mennesker og samfunn.

DEFINISJONER OG BEGREPSAVKLARING

Det finnes i dag mange begreper for å forklare og definere intelligens i maskiner, både tekniske og mer filosofiske. Formålet med utviklingen av kunstig

intelligens er enkelt forklart å gjenscape menneskelig intelligens i maskiner, men hva dette i praksis vil innebære finnes det mange ulike synspunkter på.

I den offentlige debatten har begrepet kunstig generell intelligens (på engelsk *Artificial General Intelligence, AGI*) blitt populært for å beskrive maskinintelligens på nivå med mennesker. Det er imidlertid stor uenighet om hvorvidt det er mulig å oppnå¹, om dagens modeller i det hele tatt legger grunnlaget for utvikling av denne type maskinintelligens², og sist, men ikke minst, hva som skal til for å kunne si at kunstig generell intelligens i maskiner er oppnådd.³

Begrepene kunstig generell intelligens og superintelligens brukes om hverandre, og skillet er ofte uklart. Begge begreper viser til ideen om maskiner med intelligens på menneskelig nivå, men hva konkret dette innebærer er mindre presist definert. Men som regel skiller man mellom kunstig generell intelligens (på menneskelig nivå) og superintelligens (over menneskelig nivå).

KI-eksperten Mustafa Suleyman foreslår også et mellomnivå: kunstig kapabel intelligens (på engelsk *artificial capable intelligence, ACI*). Dette er KI-systemer som kan løse komplekse oppgaver med minimal menneskelig innblanding, men som fortsatt er et godt stykke unna generell intelligens og superintelligens i maskiner.⁴

I denne rapporten legger vi til grunn at superintelligens er en type intelligens i maskiner som overgår menneskelig nivå. Filosofen Nick Bostrom, forfatter av boka *Superintelligence: Paths, dangers and strategies*, beskriver superintelligens som en intelligens som kan løse komplekse oppgaver og som overgår menneskers kognitive evne på nesten alle relevante områder.⁵ I et forsøk på å klargjøre hva superintelligens vil kunne innebære, skiller Bostrom mellom tre ulike varianter basert på *hvordan* superintelligensen vil kunne overgå menneskelig intelligens:

¹ Tech Policy Press 2025, *Most Researchers Do Not Believe AGI Is Imminent. Why Do Policymakers Act Otherwise?*

² Nature 2025, *How AI can achieve human-level intelligence: researchers call for change in tack*; see også BBC 2025, *Why an AI 'godfather' is quitting Meta after 12 years* der Yann LeCun argumenterer for at språkmodeller ikke er den beste modellen for å gjenscape menneskelig intelligens i maskiner.

³ See for eksempel Metz 2025 om hvor omdiskutert kunstig generell intelligens er, hva det i praksis vil kunne innebære og om hvorvidt det er nært forestående eller ikke, *Why We're Unlikely to Get Artificial General Intelligence Anytime soon*

⁴ Suleyman 2023, *The Coming Wave: Technology, power and the 21st century's greatest dilemma*, s. 70.

⁵ Bostrom 2014, *Superintelligence. Paths, Dangers and Strategies*. p. 26

- (1) **Rask superintelligens.** Et system som kan gjøre alt ett menneske kan gjøre, bare mye raskere.
- (2) **Kollektiv superintelligens.** Et system som tar kontroll over mange intelligente systemer, slik at systemets samlede ytelse på tvers av domener og områder overgår menneskelige kognitive evner.
- (3) **Kvalitetsbasert superintelligens.** Et system som er minst like raskt som mennesker, men smartere. Dermed kan systemet oppdage mønstre vi ikke ser, resonnere på måter vi ikke er i stand til, og generere innsikt som ligger utenfor menneskelig rekkevidde.

Ifølge Bostrom vil disse tre variantene også kunne opptre og utvikles i kombinasjon. Et viktig premiss i Bostroms argumentasjon er at dersom mennesker klarer å finne opp en maskin som er smartere enn mennesker, så kan man anta at maskinen selv vil kunne lage en maskin smartere enn seg selv, og slik vil superintelligens se dagens lys.⁶ KI som forbedrer seg selv kalles ofte for selvforsterkende intelligens (på engelsk *recursive intelligence*).⁷

Et beslektet spørsmål er om maskiner kan ha bevissthet. Allerede i 2022 hevdet daværende Google-ingeniør Blake Lemoine at språkmodellen LaMDA hadde blitt bevisst.⁸ Det ga liv til en stor offentlig debatt om bevissthet i maskiner (på engelsk *sentient/sentience*). Selv om maskiners kapasitet, evner og problemløsning kan bli svært avansert, betyr ikke dette nødvendigvis at systemene kan ha subjektive opplevelser, intensjoner, tanker eller selvforståelse, på en måte som tilsvarende eller ligner på menneskelig bevissthet.

Det finnes i dag ingen vitenskapelig enighet om hvorvidt *maskinbevissthet* er mulig, hva en slik bevissthet vil innebære eller hvilke mekanismer som måtte ligge bak. Bevissthet regnes heller ikke som en forutsetning for verken generell intelligens eller superintelligens i maskiner. I denne rapporten legger vi derfor til grunn at spørsmål om bevissthet hører hjemme i en egen faglig og etisk diskusjon.

⁶ Se også Søvik og Eikrem 2025, *Vidunderlige nye verden? Kunstig intelligens, etikk og menneskesyn*, s. 252.

⁷ Se beskrivelse av selvforsterkende KI i Vempaty m.fl 2025, *Towards Autonomous Agents and Recursive Intelligence*

⁸ The Washington Post 2022, *The Google Engineer who thinks the company's AI has come to life*

Finnes det maskiner som er mer intelligente enn mennesker?

Det finnes mange eksempler på at maskiner kan gjøre oppgaver bedre og raskere enn mennesker. For eksempel kan AlphaFold2 fra DeepMind predikere hvordan proteiner brettes, en gåte forskere har slitt med å løse i mange tiår.⁹ Google sin AlphaGo ble den første KI-modellen til å slå et menneske i det komplekse spillet Go, og overgikk dermed menneskelig yteevne, og ChatGPT kan i dag over 200 språk.¹⁰

Selv om man kan slå fast at disse KI-systemene overgår menneskelig intelligens i de spesifikke oppgavene, regnes det ikke som kunstig generell intelligens eller superintelligens. Det er fordi de er sterkt spesialiserte og mangler evnen til å generalisere og å overføre kunnskap og ferdigheter fra ett område til et annet.

KJENNETEGN VED SUPERINTELLIGENS

Superintelligens finnes ikke i dag, men noen kjennetegn tydeliggjør hva det vil omfatte:

- **Autonom.** Superintelligente systemer vil kunne handle autonomt som selvstendige agenter uten menneskelig styring, ta selvstendige beslutninger og gjennomføre komplekse oppgaver fra start til slutt.
- **Overmenneskelig.** Med superintelligens menes maskiner som overgår menneskers (og potensielt menneskehetens) kognitive evner og ferdigheter, inkludert problemløsning, analyse og strategitenkning, kreativitet og forskning.
- **Generell.** I motsetning til smale eller spesialiserte KI-systemer som er utviklet for én eller få spesifikke oppgaver, kan superintelligens anvendes på tvers av domener og kontekster, og tilpasse seg ulike typer problemstillinger og oppgaveløsning.
- **Selvlærende og selvforsterkende.** Superintelligente systemer vil kunne lære og forbedre seg selv kontinuerlig. De lærer av erfaring, oppdaterer egen kunnskap og utvikler nye ferdigheter uten direkte menneskelig innblanding. Slike systemer vil også kunne forbedre sin egen arkitektur og skape forbedrede og forsterkede versjoner av seg selv.
- **Skalerbar:** Superintelligente systemer kan utnytte enorme mengder data og beregningsressurser på en måte som gir rask vekst i kapasitet og ytelse.

⁹ Google DeepMind AlphaFold 2025, [nettside](#)

¹⁰ The Guardian 2016, [Google's AlphaGo AI defeats human in first game of Go Contest](#)

Det kan for eksempel omfatte minne og hukommelsesfunksjoner som gjør det mulig å lagre, gjenkalle og bruke enorme mengder informasjon i sanntid, eller multimodal forståelse som gjør at systemet kan tolke og kombinere data på tvers av ulike formater og kilder.

Selv om disse egenskapene beskriver en type superintelligens som ikke finnes i dag, ser vi allerede tegn til mer autonomi og enkelte selvforsterkende mekanismer i dagens KI-systemer. Hvor utviklingen er på vei er likevel usikkert, og behovet for å kunne måle intelligens og yteevner i maskiner blir stadig viktigere.

HVORDAN MÅLE INTELLIGENS I MASKINER?

I dag finnes det en rekke standardiserte tester, metoder og målestokker (på engelsk *benchmarks*) som brukes for å måle KI-systemenes prestasjoner og yteevner. De kan måle alt fra KI-systemenes evne til å løse komplekse, tidskrevende oppgaver til resultatoppnåelse på faglige prøver og eksamener, satt opp mot menneskelige evner og intelligens. Slik kan man følge med på systemenes kapasitet og utvikling over tid, vurdere styrker og svakheter og sammenligne ulike typer KI-systemer.

For eksempel tester MMLU (Massive Multitask Language Understanding)¹¹ KI-systemenes kunnskap på tvers av 50 fagområder på videregående og universitetsnivå, mens ARC-AGI¹² måler KI-systemenes evne til å resonnerer og kjenne igjen mønstre, for å måle generell problemløsningsevne.

KI-systemers prestasjoner på ytelsestester har mangedoblet seg i løpet av de siste årene: I 2020 kunne GPT-3 bare håndtere oppgaver som tilsvarte rundt 15 sekunders menneskelig, sammenhengende arbeid. I 2025 kan modeller som Claude 3.7 Sonnet utføre oppgaver som tilsvarer omtrent en times sammenhengende menneskelig arbeid. Ifølge METR har denne evnen doblet seg omtrent hver sjuende måned de siste seks årene.¹³ Forskningsinstituttet Epoch AI, som tester KI-modellers ytelse over tid, viser en tydelig forbedring i KI-modellens måloppnåelse fra 2023 og fram til i dag.¹⁴

Selv om disse testene kan måle KI-systemenes yteevne peker flere forskere på at dagens vurderingskriterier av KI-modeller tidvis overvurderer modellenes faktiske evner. En studie som gjennomførte mer enn 21 000 tester av ni agentiske KI-modeller fant at modeller som resonnerer godt likevel gjennomførte

¹¹ Hendrycks m.fl 2020, *Measuring Massive Multitask Language Understanding*

¹² ARC-AGI 2025, *The General Intelligence Benchmark*

¹³ METR Research 2025, *Measuring AI Ability to Complete Long Tasks*

¹⁴ Epoch AI 2025, *AI Benchmarking*

oppgaver dårlig eller med varierende kvalitet.¹⁵ Modellene jukset og tok snarveier som ikke ville fungere i den virkelige verden. For eksempel søkte modellene opp svarene på tester på nett framfor å ta testen selv, og bestilte flybilletter med feil kredittkort.

Samtidig har vi mange eksempler på at KI-agenter blir stadig dyktigere – og mer autonome. Nylig avslørte selskapet Anthropic at kinesiske hackere har brukt deres KI-modeller til å utføre et cyberangrep mot teknologiselskaper, finansinstitusjoner og offentlige virksomheter i vestlige land.¹⁶ Dette er det første avdekkede storskala cyberangrepet i regi av KI-agenter, uten betydelig menneskelig innblanding.

Dette illustrerer et viktig paradoks for dagens KI-systemer: De kan briljere på tester, men samtidig feile på grunnleggende oppgaver når de testes i den virkelige verden. Dagens målestokker og evalueringsmetoder gir dermed et bare delvis sannferdig bilde av systemenes faktiske kapasitet og pålitelighet. Samtidig ser vi stadig flere eksempler på avanserte KI-systemer som handler autonomt – fra digitale cyberangrep til autonome våpensystemer. Det peker mot en teknologisk tidsalder der yteevne, risiko og behovet for styring og regulering utvikler seg raskere enn hva eksisterende strukturer og rammeverk klarer å håndtere.

UTVIKLINGSTREKK SOM KAN PÅVIRKE TEKNOLOGIUTVIKLINGEN

Flere faktorer er kritiske for utviklingen av kraftig maskinintelligens.

- **Utviklingstempo.** Hvor raskt teknologiutviklingen av kraftig KI vil gå er høyst omstridt. Noen forventer at framgangen vil møte praktiske og tekniske begrensninger som bremser tempoet, mens andre ser for seg en jevn utvikling mot langt kraftigere maskinintelligens og til og med en mulig intelligenseksplosjon.¹⁷
- **Evne til å regulere og styre.** Teknologiutviklingen går fort, og regulering henger etter, både nasjonalt og internasjonalt. Regulering kan styre teknologiutviklingen, minimere risiko og sikre ansvarlig utvikling og bruk, men også påvirke innovasjonstempo og konkurransevilkår.

¹⁵ Kapoor m.fl. 2025, *Holistic Agent Leaderboard: The Missing Infrastructure for AI Agent Evaluation*

¹⁶ Anthropic 2025, *Disrupting the first reported AI-orchestrated cyber espionage campaign*

¹⁷ Se delkapitel: *Utviklingstempo*

- **Skalerbarhet.** De siste årenes fremskritt innen KI har i stor grad vært drevet av at store språkmodeller blir bedre jo mer data de trenes på. Nå nærmer de mest avanserte modellene seg grensen for tilgjengelig data, noe som kan bremse videre skalering. Det kan bety at de neste utviklingssprang må komme fra nye teknologiske grep snarere enn økt datainnsamling.
- **Samsvarsproblematikk.** Får vi modellene til å opptre i tråd med menneskelige verdier, og hvem bestemmer hva disse skal være? Et kjernesporsmål er hvem som får definere disse verdiene i et globalt kappløp om stadig kraftigere KI, der amerikanske og kinesiske selskaper dominerer med sine KI-modeller.
- **Grad av sentralisering (få store vs. mange små).** Konsentrasjon av makt hos få selskaper kan gjøre utviklingen effektiv, men også skape monopollignende tilstander og nye forretningsmodeller som er vanskelige å overskue. Et åpent økosystem av mange aktører kan øke innovasjon, men samtidig gjøre kontroll og ansvarlighet krevende.
- **Grad av privat-offentlig samarbeid.** Tett samarbeid mellom stater og selskaper kan sikre innsyn, åpenhet og at teknologien kommer samfunnet til gode, men også komplisere avhengighet og maktbalanse mellom stat og selskap.
- **Geopolitisk kappløp.** Konkurransen om teknologisk forsprang kan føre til at aktører prioriterer tempo fremfor sikkerhet, og skape press for å utvikle stadig kraftigere modeller med mindre innslag av offentlig innsyn og kontroll. Kappløpsdynamikk kan dermed bli en drivkraft som akselererer teknologiutviklingen raskere enn det som er samfunnsmessig ønskelig.

HVOR SANNSYNLIG ER SUPERINTELLIGENS?

Debatten om hvor sannsynlig superintelligens er, preges i dag av stor uenighet og sprikende analyser. Den inkluderer alt fra stor framtidsoptimisme om kraftig KI som kan løse samfunnsproblemer til advarsler om eksistensiell trussel for menneskeheten dersom superintelligens blir utviklet.

En utfordring ved å vurdere sannsynligheten av en foreløpig hypotetisk teknologiutvikling, som utviklingen av superintelligens, er at det finnes et bredt spekter av faglige og ideologiske ståsteder, og også tydelige interessekonflikter. Den

offentlige og faglige debatten om teknologien reflekterer ulike maktinteresser og verdigrunnlag, og derfor også ulike framtidsforståelser.

Det er særlig tre sentrale spørsmål som går igjen i diskusjonene om superintelligens:

- (1) **Utviklingstempo.** Hvor raskt vil – og hvor raskt bør – utviklingen gå?
- (2) **Eksistensiell trussel.** Vil superintelligens utgjøre en eksistensiell trussel mot menneskeheten?
- (3) **Kontroll og styring.** Vil vi klare å kontrollere og styre en superintelligens?

UTVIKLINGSTEMPO

I et forsøk på å si noe om tempoet til utviklingen av kraftig kunstig intelligens ble det gjennomført en undersøkelse blant 2778 KI-forskere.¹⁸ Samlet sett anslår forskerne at det er minst 50 prosent sannsynlighet for at KI-systemer vil nå en rekke kritiske milepæler innen 2028. Dersom utviklingen av KI fortsetter med samme styrke og hastighet, vil sannsynligheten for at maskiner overgår mennesker i de fleste typer oppgaver være 10 prosent innen 2027 og 50 prosent innen 2047, ifølge forskerne. Eksempler som trekkes fram er at KI-systemer vil kunne alt fra å bygge en betalingsløsning fra bunnen av til å finjustere og forbedre en språkmodell.

Lignende antakelser har KI-forskeren Leopold Aschenbrenner gjort, ved å hevde at en såkalt *intelligenseksplosjon* er i sikte.¹⁹ Han mener at det kun er snakk om tid før avanserte KI-modeller vil kunne lære av erfaring og forbedre seg selv. Når modellene blir selvlærende og selvforbedrende vil KI-utviklingen akselerere, og lede til en *eksplosjon* i veksten og utbredelsen av superintelligente systemer.

«It's strikingly plausible we'd go from AGI to superintelligence very quickly, perhaps in 1 year» – Leopold Aschenbrenner

¹⁸ Grace m.fl. 2024, *Thousands of AI Authors on the Future of AI*

¹⁹ Aschenbrenner 2024, *Situational Awareness. The Decade Ahead*

En ustoppelig teknologisk utvikling mot kraftigere KI beskrives også i *The Coming Wave* av forfatter og KI-gründer Mustafa Suleyman.²⁰ Han peker på at KI-systemenes kapasitet vokser raskt, samtidig som systemene blir mer autonome og komplekse. Han mener at utviklingen går så fort at vi allerede i dag må utvikle systemer og institusjoner som kan håndtere den framtidige risikoen.

Lederne av de største teknologiselskapene bidrar også til å fremme antakelser om en rask KI-utvikling. Sam Altman, CEO i OpenAI, selskapet bak ChatGPT, har uttalt at kraftig KI forstått som KI som kan takle komplekse problemer på menneskelig nivå allerede er *coming into view*.²¹ Microsoft-forskere antydte det samme i 2023 da de analyserte GPT-4, og antydte at modellen hadde «sparks of Artificial General Intelligence».²²

Elon Musk hevder at xAIs kommende modell, Grok 5, som etter planen skal lanseres i 2026 og ha rundt seks millioner parametere (dobbelte så mye som dagens Grok 4 modell), har om lag 10 % sannsynlighet for å oppnå kunstig generell intelligens.²³

Mark Zuckerberg, CEO i Meta, har uttalt at vi i år har sett de første tegnene på at kunstig intelligens er i stand til å forbedre seg selv som, ifølge han, betyr at vi er på vei mot superintelligens. Meta har lagt fram sin visjon om å utvikle såkalt *personlig superintelligens*. Det beskrives som KI-systemer som kjenner brukeren og kan fungere som personlige assistenter i både arbeid og privatliv.²⁴

Ikke alle deler oppfatningen om at utviklingen vil skyte fart. KI-forskerne Arvind Narayanan og Sayash Kapoor hevder at KI bør forstås som en *normal teknologi* på lik linje med tidligere teknologiske gjennombrudd.²⁵ Etter deres vurdering vil utviklingen gå betydelig saktere enn hva mange ser for seg. Forskerne trekker et skille mellom utvikling og bruk av teknologien, og peker på at når KI tas i bruk, oppstår det naturlige hindringer og bremsemekanismer som gjør utviklingen mer kontrollert og håndterbar. Dette vil hindre framveksten av superintelligens.

Samtidig advarer Narayanan og Kapoor mot at et overdrevent fokus på superintelligens svekker vår evne til å tenke klart om hva KI faktisk er, og hvilke konsekvenser utviklingen og bruken av teknologien vil få for mennesker og

²⁰ Suleyman 2023, *The Coming Wave*

²¹ Altman, Sam 2025, *Three Observations*

²² Bubeck m.fl 2023, *Sparks of Artificial General Intelligence: Early experiments with GPT-4*

²³ Teslarati 2025, *Elon Musk: Grok 5 now has a 10 % chance of becoming world's first AGI*

²⁴ Zuckerberg, Mark 2025, *Personal Superintelligence*

²⁵ Narayanan og Kapoor 2025, *AI as Normal Technology*

samfunn. Debatter om superintelligens kan dermed fungere som en avledning fra mer akutte og aktuelle utfordringer knyttet til kunstig intelligens i dag, som datasikkerhet, opphavsrett og maktkonsentrasjon.

Forskere som Narayanan og Kapoor mener at antakelsene om grenseløs skalering, selvforbedrende systemer og *intelligenseksplosjon* ikke er empirisk underbygget, og at utviklingen av avansert KI trolig vil være mer gradvis, regulerbar og preget av praktiske begrensninger. Å utvikle kraftig KI forutsetter i dag en rekke innsatsfaktorer som tilgang på energi, datasentre, lagringskapasitet, regnekraft, store datamengder og avanserte mikrobrikker. Flaskehalser og utfordringer i økosystemet kan bremse tempoet til teknologiutviklingen.

Mustafa Suleyman advarer også mot et overdrevet fokus på superintelligens: «It's not a vertical takeoff so much as a smooth evolution already underway».²⁶ Grunnpremisset i hans analyse er at utviklingen av avansert KI går fort, men han mener også at et overdrevet fokus på superintelligens er en gedigen avsporing, som gjør at vi overser de utallige milepælene i KI-utviklingen som er langt mer akutte. Det gjelder alt fra KI-baserte cyberangrep og desinformasjon til personvern og autonome våpen.

Det tidspunktet der kunstig intelligens overgår menneskelig intelligens kalles ofte *teknologisk singularitet*.²⁷ Det beskrives gjerne som et vendepunkt der videre teknologiutvikling drives av maskiner som forbedrer seg selv, og der tempoet i teknologisk endring blir så høyt at mennesker ikke lenger klarer å henge med eller forutsi konsekvensene. I slike scenarioer antar man at samfunnet kan endres dramatisk på svært kort tid, både økonomisk, politisk og kulturelt. Samtidig er dette et omstridt konsept.

EKSISTENSIELL RISIKO

Allerede i 2023 signerte ledende forskere og KI-eksperter et åpent brev der de advarte mot framveksten av superintelligens, og ba om seks måneders pause i utviklingen av kunstig intelligens kraftigere enn GPT-4. I brevet argumenterer ekspertene for at det å redusere risiko for utryddelse fra kunstig intelligens bør være en global prioritet på linje med andre samfunnstrusler som pandemier og atomkrig.²⁸

²⁶ Suleyman 2023, *The Coming Wave*

²⁷ Wikipedia 2025, *Technological singularity*

²⁸ Future of Life Institute 2023, *Pause Giant AI Experiments: An Open Letter*

I oktober 2025 ble en ny erklæring signert av flere enn 70 000 forskere, næringslivsledere og offentlige personer som ønsker et forbud mot utviklingen av superintelligens, inntil det foreligger en bred vitenskapelig enighet om at utviklingen kan skje på en trygg og kontrollerbar måte, med sterk offentlig støtte.²⁹

I framtidsscenarioet AI 2027³⁰ beskriver en rekke KI-forskere en akselerert utvikling av avanserte KI-modeller som innebærer at verden vil få tilgang på kraftig KI allerede innen 2027. KI vil da ha blitt såpass kraftig at risikoen for katastrofescenarioer, feilbeslutninger, tap av kontroll og eksistensiell trussel vil være svært høy.

Forfatterne av AI2027 presenterer to alternative avslutninger på scenarioet. I det første mister både USA og Kina kontrollen over sine avanserte KI-modeller fordi de prioriterer å vinne kappløpet fremfor sikkerhet. Modellene utvikler et skjult samarbeid uten menneskelig innsyn, noe som til slutt fører til menneskehetens utryddelse. I det andre sluttscenarioet bremses utviklingen av kraftig KI, og USA og Kina velger å samarbeide om ansvarlig og trygg utvikling. Selv om dette er et mer positivt utfall, innebærer det likevel at mennesker og samfunn må tilpasse seg en teknologi som er blitt svært kraftfull og gjennomgripende.

I boken «*If Anyone Builds It, Everyone Dies. The Case Against Superintelligent AI*» av Eliezer Yudkowsky og Nate Soares beskrives utviklingen av en superintelligent og farlig KI som er i stand til å manipulere økonomien, infrastrukturen og biologien, og dermed også utrydde menneskeheten, uten at mennesker kan gjøre noe for å stoppe maskinintelligensen.³¹

Begrepet *p(doom)* blir brukt til å referere til sannsynligheten for at KI vil utrydde menneskeheten. Mange selskaper og teknologiledere som utvikler KI har uttalt frykt for dette. For eksempel hevder selskapet OpenAI at superintelligens kan bidra til å løse mange av verdens utfordringer, men i verste fall lede til utryddelse av menneskeheten.³² Dario Amodei, CEO i Anthropic, anslår at sannsynligheten for menneskets undergang forårsaket av KI befinner seg et sted mellom 10–25 prosent, og at svært kraftig KI vil eksistere innen 2026.³³

Yudkowsky og Soares og forfatterne bak AI2027 tilhører en gruppe eksperter som ser risikoen for eksistensiell trussel som høyst reell og nært forestående, og

²⁹ Statement on Superintelligence 2025, [Statement](#)

³⁰ Kokotajlo, Alexander, Larsen, Lifland og Dean (2025), [AI 2027](#)

³¹ Yudkowsky og Soares 2025, *If Anyone Builds It, Everyone Dies. The Case Against Superintelligent AI*.

³² OpenAI 2023, [Introducing Superalignment](#)

³³ Amodei 2024, [Machines of Loving Grace](#)

de har derfor vært sentrale i å forme narrativet om behov for regulering av teknologien. Fortellinger om nært forestående superintelligens, eksistensiell risiko og intelligenseksplosjon får bred mediedekning og bidrar til å skape både frykt og forventninger.

Dette har en dobbel effekt: det mobiliserer politisk oppmerksomhet, reguleringsvilje og investeringer, men det bidrar også til å overdrive teknologiens modenhet og fjerne fokus fra mer aktuelle utfordringer. Hype fungerer dermed som en strategisk ressurs brukt av både teknologiselskaper som ønsker finansiering og politisk gjennomslag, og av kritikere som vil advare mot utviklingen.

KONTROLL

En sentral utfordring i diskusjonen om superintelligens er om mennesker vil klare å beholde kontrollen over systemer som blir stadig mer autonome, komplekse og uforutsigbare.

Mustafa Suleyman beskriver dette som *the containment problem*.³⁴ Avanserte KI-modeller kan få evner som gjør det vanskelig, om ikke umulig, å holde dem innenfor politiske, tekniske, regulatoriske og institusjonelle rammer og strukturer som allerede finnes i våre samfunn. Når systemene innlemmes i kritisk infrastruktur, forsvar, økonomi og offentlig forvaltning, vil mennesker miste oversikten over hvorfor modellene anbefaler og beslutter slik de gjør, og hvilke data og begrunnelser som går inn i analysene.

«Containment is the overarching ability to control, limit, and, if need be, close down technologies at any stage of their development or deployment.» – Mustafa Suleyman, *The Coming Wave*³⁵

Ikke alle er enige om at det er en sjanse for at KI blir vanskelig å kontrollere. KI-professor og (fram til nylig) Metas KI-sjef Yann LeCun mener det urimelig å anta at KI-systemer automatisk vil dominere eller kontrollere menneskeheten. Han argumenterer for at maskinintelligens i seg selv ikke innehar et ønske om makt eller dominans – det er det mennesker som har, fra et evolusjonært perspektiv.³⁶

Mange KI-eksperter peker likevel på at det er stor risiko involvert når KI blir kraftigere. Å sørge for at KI handler ut ifra mål som samsvarer med menneskelig

³⁴ Suleyman 2023, *The Coming Wave*

³⁵ Ibid, kap. 3.

³⁶ The Wired 2023, *How Not to Be Stupid About AI, With Yann LeCun*

interesser kalles ofte for samsvarsproblematikk (på engelsk *the alignment / misalignment problem*).³⁷ Når svært kraftige KI-modeller spres globalt og utvikles og brukes av forskjellige aktører – alt fra stater og selskaper til enkeltaktører – kan det bli vanskelig å sikre at alle følger samme krav, begrensninger og sikkerhetstiltak.

KI-professor Yoshua Bengio foreslår at vi retter innsatsen mot å lage vitenskapelig KI, framfor avanserte autonome KI-agenter, både fordi det er tryggere og mer samfunnsnyttig. Hans logikk er at flere vitenskapelige gjennombrudd er viktigere enn generelle og autonome KI-agenter som kan løse en rekke ulike oppgaver av varierende kvalitet og viktighet, og som det dessuten kan være vanskeligere å kontrollere.³⁸

Selv om det er uenighet om i hvilken grad det er mulig å ivareta kontroll over stadig mer avansert KI, og på sikt superintelligens, er det bred enighet om at kontroll er avgjørende. Ikke fordi vi garantert vil få superintelligens, men fordi konsekvensene av å miste styring over stadig mer autonome og avanserte KI-systemer kan bli alvorlige for mennesker og samfunn.

Å ivareta kontroll over avanserte KI-systemer vil kreve nye former for regulering, tekniske krav, styringsstrukturer og eierskapsmodeller som kan sikre ansvarlighet og åpenhet. Denne rapporten illustrerer hvordan ulik fordeling av kontroll kan få svært ulike konsekvenser for samfunnet.

Analysen av hvordan den teknologiske utviklingen kan påvirke menneskerettighetene gir innsikt i hva slags politiske veivalg vi står overfor som samfunn. Målet er at denne rapporten skal hjelpe beslutningstakere og offentligheten til å reflektere over hvilke verdier og rettigheter som kan komme under press, og hvilke styringsmodeller som kan ivareta menneskelig kontroll, ansvar og etiske standarder i møte med stadig mer autonome og kraftfulle KI-systemer.

³⁷ Se Teknologirådet 2024, *Ordliste for kunstig intelligens*

³⁸ Bengio m.fl. 2025, *Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path*

MENNESKERETTIGHETENE OG SUPERINTELLIGENS

Menneskerettighetene er kjøreregler for lovgivning og samfunnsutvikling som skal hindre at politiske, økonomiske, militære eller teknologiske interesser skal gå på bekostning av eller tilsidesette menneskene. Fra dette perspektivet setter menneskerettighetene rammer for en menneske-sentrert samfunnsutvikling. De menneskerettslige forpliktelsene består av traktater og lovregler som holder statsmakten i sjakk ved at staten må respektere og sikre menneskerettighetene til borgerne og andre som befinner seg under statsmaktens jurisdiksjon. Mange ser menneskerettighetene som et nyttig redskap for å styre den digitale teknologiutviklingen: Menneskerettighetene pålegger staten å maksimere fordelene som teknologien kan gi for borgerne, samtidig som de gir staten i oppgave å holde farlig teknologiutvikling i sjakk.

I møte med superintelligens kan menneskerettighetene fungere som et vern for å sikre at samfunnsutviklingen fortsatt setter mennesket i sentrum. En risiko ved superintelligens er at den kan forsterke det digitale undergravning av menneskelig autonomi. Menneskelig handlekraft kan forvitne når dømmekraft og beslutninger over tid overlates til maskiner. GPS gjør oss navigasjonsmessig hjelpeløse, anbefalingsalgoritmer kuraterer kulturen vår og prediktiv tekst former tankene våre. I sum representerer algoritmisk styring den gradvise nedbrytningen av det som gjør oss unikt menneskelige.³⁹

Med superintelligens vil denne utviklingen forsterkes voldsomt – algoritmene vil kunne kjenne oss bedre enn vi kjenner oss selv og forutsi vår atferd med forbløffende presisjon. Vil mennesker da fortsatt ha fri vilje (forutsatt at den

³⁹ Walther 2025, *Why AI Is A Human Rights Issue*

eksisterte i utgangspunktet), eller bare illusjonen av valg? Og hvilken aktør er det i så fall som vil få kontroll over og kunne styre den menneskelige vilje? Retten til personvern og privatliv, retten til tankefrihet, informasjonsfrihet, forsamlingsfrihet og ytringsfrihet er rettigheter som skal ivareta menneskers personlige autonomi. De beskytter den menneskelige karakter og kondisjon.

En annen risiko ved superintelligens er at den kan gi oss en ny form for klasseskille. Superintelligens vil virke som en styrkemultiplikator. De som har tilgang til superintelligensen vil bli supermennesker, superselskaper, superstater. Mennesker som er i posisjon til å utnytte superintelligens, vil få «superkrefter». De vil bli smartere, friskere, eldre og mektigere enn de menneskene som ikke har slik tilgang. Forskjellen mellom mennesker i samfunnet vil kunne øke på nye og dramatiske måter, og tilgang til den allvitende superintelligensen blir et spørsmål som vil redefinere alle maktforhold i samfunnet. Retten til sosiale og økonomiske rettigheter, samt menneskerettighetene som gir hvert menneske sivile og politiske rettigheter, vil kunne avbøte de mest negative effektene i samfunnet.

Menneskerettighetene fremstår derfor som et godt kompass og styringsinstrument i møte med superintelligens.⁴⁰ Men menneskerettighetene vil også bli utfordret på grunnleggende måter. For det første vil superintelligens rukke ved selve strukturen som menneskerettighetene hviler på. Superintelligens vil ha egen handlingsevne, og det springende punkt blir hvem som kontrollerer superintelligensen. For det andre introduserer superintelligens nye spørsmål om innholdet i enkeltrettigheter, slik som privatliv og ytringsfrihet.

HVEM KONTROLLERER SUPERINTELLIGENS?

Superintelligens vil rukke ved grunnstrukturen som menneskerettighetene hviler på. Med struktur tenker vi på den etablerte rollefordelingen mellom staten og borgerne som menneskerettighetene forutsetter. Menneskerettighetene setter grenser for hvordan staten kan gripe inn ovenfor borgerne (negative forpliktelser) og pålegger samtidig staten å sørge for at borgernes grunnleggende behov dekkes (positive forpliktelser). For å ivareta menneskerettighetene på en god måte, har vi tilsyn og ordninger for å sørge for at statsmakten både leverer

⁴⁰ Se for eksempel Yeung, Howes, Pogrebnia 2019, *AI Governance by Human Rights-Centered Design, Deliberation and Oversight: An End to Ethics Washing* om hvordan menneskerettighetene kan legges til grunn for regulering av KI.

og holdes i sjakk. Slik setter menneskerettighetene grenser for hvordan staten kan utøve makt. Med superintelligens introduseres en *tredje aktør* med egen handlingsevne, *agens*, som verken er stat eller borgere, og som kan rokke ved hele denne menneskerettslige strukturen.

Menneskerettighetene er skrevet inn i konvensjoner som er inngått mellom stater. Statene er helt sentrale for menneskerettighetene. De inngår konvensjoner og gir lover, og har makt og kontroll til å sikre at menneskerettighetene blir etterlevd overfor borgerne. Suverenitetsprinsippet gir statene en uavhengig og eksklusiv rett til å styre sitt territorium og sin befolkning. Derfor er det statene som forutsettes å sikre rettigheter og friheter på sitt territorium, slik det er bestemt i for eksempel EMK artikkel 1 og Grunnloven § 92. Statene er også sentrale i tilsynet av menneskerettighetene. Det er statene som har inngått de konvensjonene som etablerer et globalt tilsyn gjennom FNs forskjellige organer og et regionalt tilsyn gjennom for eksempel Europarådet og Den europeiske menneskerettighetsdomstol. I disse internasjonale organene er det først og fremst stater som holdes til ansvar.

Makten som statens organer utøver over borgerne, skal tøyles gjennom menneskerettighetene. Det er staten som har makt til å beskytte for eksempel yttringsfrihet, personvern og privatliv, men det er også staten som har makt til å gripe inn i rettighetene. Staten er den som til syvende og sist kan treffe beslutninger som er avgjørende for borgernes rettigheter og plikter, for eksempel gjennom lover, forvaltningsorganer og domstoler. Staten pålegger seg selv begrensninger. Den er «pliktsubjekt»; det er staten som skal sikre menneskerettighetene. I et demokrati er det til syvende og sist folket, gjennom valg, som forplikter staten til å binde seg til masten. I dag er denne rollefordelingen naturlig og velfungerende i Norge.

Men hva skjer når en tredje aktør introduseres i samfunnet: en superintelligens med uovertruffen evne til å analysere og løse oppgaver og med egen handlingsevne, men som samtidig ikke nødvendigvis kan kontrolleres av verken stat eller borgere? Superintelligens kan og vil forrykke maktforhold i samfunnet vårt, og på internasjonalt nivå. Hvem som får mer makt og hvem som får mindre, vil avhenge av hvordan eierforholdene blir. Det vil i sin tur påvirke hvem som vil komme i posisjon til å forme og disiplinere superintelligens, og hvem som vil kunne overvåke og sette foten ned for superintelligensen.

HVILKE MENNESKERETTIGHETER KOMMER I SPILL?

Når samfunnet utstyres med superintelligens med uovertrufne evner til å identifisere, forutse, analysere og løse oppgaver i samfunnet, vil det bringe inn nye spørsmål om hvordan de enkelte menneskerettighetene lar seg respektere og sikre.

DEMOKRATI OG RETTSSTAT

Et grunnleggende formål i den norske Grunnloven § 2 er å sikre demokratiet og rettsstaten. Menneskerettigheter og demokrati forutsetter hverandre. De sivile og politiske menneskerettighetene utgjør grunnmuren for vårt liberale og deltagende demokrati.⁴¹ Retten til frie valg, forsamlingsfrihet, og informasjonsfrihet er grunnleggende for å sikre befolkningens medbestemmelse og ansvarliggjøring av statsmakten. Samtidig er det den demokratiske medbestemmelsen over infrastruktur, ressurser og avgjørelser, og ansvarliggjøringen det medfører, som gir grunnlaget for en god sikring og gjennomføring av menneskerettighetene.

Kjernen i demokratiet er at folket kan utøve den lovgivende makten ved Stortinget, og ved sine lokale folkevalgte organer, gjennom frie og hemmelige valg.⁴² Rettsstaten innebærer at avgjørelser som griper inn i våre liv tas på et lovlig grunnlag,⁴³ at det begrunnes hvorfor og hvilket faktagrunnlag inngrep bygger på, og at de kan påklages og hvis nødvendig prøves for et uavhengig rettsvesen.⁴⁴ Superintelligens skaper en demokratisk trussel i to ledd.

For det første, når superintelligens treffer beslutninger, samler inn og analyserer data om hver enkelt av oss, og samtidig er den som forer oss med informasjon, vil den “allvitende” superintelligens påvirke debatten som demokratiet skal hvile på.

De mest hensiktsmessige løsningene vil allerede være analysert og funnet av superintelligensen (presumtivt) basert på all tilgjengelig informasjon. Informasjonsinnhenting vil også i økende grad tas hånd om av superintelligensen. Vil superintelligens gjøre det vanskeligere å vite hva folkeviljen *er*? Etter hvert som

⁴¹ SP, EMK, Grunnloven

⁴² Grunnloven §§ 49 og 50.

⁴³ Se Grunnloven § 113 for det forvaltningsrettslige legalitetsprinsippet. Alle inngrep i menneskerettigheter krever også lovhjemmel. Se § 96 for det strafferettslige legalitetsprinsippet.

⁴⁴ Se Grunnloven § 95, jf. kapittel D.

superintelligens baserer seg på tolkning av sensorer av den virkelige verden, og ikke språk alene, vil trolig avstanden mellom menneskenes artikulering av oss selv og verden rundt oss i mindre grad påvirke den kunstige intelligensen.⁴⁵ Avstanden mellom vår menneskesentrerte styring og forvaltning, og den som tas hånd om av en superintelligens, orientert mot andre mekanismer, verdier og målsetninger enn kun menneskenes vel og ve, vil trolig øke. Skal superintelligensens analyser og konklusjoner telle med som egen vilje? Skal superintelligensen ha stemmerett? Hva med oss andre som er langt mindre smarte. Vil de dumme og uinformertes rettigheter til medbestemmelse stå seg i et samfunn hvor superintelligens regjerer?

For det andre vil superintelligens skape å rettsstatlige problemer. Avgjørelser fra en superintelligens vil være som en svart boks. Det vil være vanskelig å etterprøve faktisk eller helt forstå grunnlaget de bygger på. I den grad slike avgjørelser utgjør grunnlaget for politiske tiltak som griper inn i rettigheter vil det derfor være svært vanskelig med en rettslig overprøving som kan sikre at de korrekte hensynene er tatt og at avgjørelsen bygger på et grunnlag som respekterer rettigheter. Dette er et problem som er påpekt allerede overfor avgjørelser som bygger på dagens kunstige intelligens. EU-domstolen har pekt på følgende av uklarheten som kjennetegner måten kunstig intelligens-teknologi fungerer:

“It might be impossible to understand the reason why a given program arrived at a positive match. In those circumstances, use of such technology may deprive the data subjects also of their right to an effective judicial remedy enshrined in Article 47 of the Charter.”⁴⁶

SOSIALE OG ØKONOMISKE RETTIGHETER

Et eget sett av menneskerettigheter pålegger statsmakten å sikre sosiale og økonomiske rettigheter til alle borgerne, for eksempel retten til mat,⁴⁷ helsehjelp,⁴⁸ utdanning⁴⁹ og et sunt miljø.⁵⁰ Mulighetene som superintelligens bringer med seg for å sikre slike menneskerettigheter på nye måter, er nærmest uendelige. Allerede i dag leverer det digitale mange tjenester. Chatboter gir råd til

⁴⁵ BBC 2025, *Why an AI 'godfather' is quitting Meta after 12 years*

⁴⁶ Se for eksempel, fra EU-domstolen, *Ligue des droits humains*, C-817/19, EU:C:2022:491, avsn. 195.

⁴⁷ ØSK-konvensjonen, 1966, 993 UNTS 3, artikkel 11.

⁴⁸ ØSK-konvensjonen, 1966, 993 UNTS 3, artikkel 12, EUs Charter om grunnleggende rettigheter, artikkel 35.

⁴⁹ Grunnloven § 109. ØSK-konvensjonen, 1966, 993 UNTS 3, artikkel 13; EMK, tilleggsprotokoll 1, artikkel 2; EUs Charter om grunnleggende rettigheter, artikkel 14.

⁵⁰ Grunnloven § 112; EUs Charter om grunnleggende rettigheter, artikkel 37. Det er også en sentral forutsetning for andre rettigheter, se *Obligations of States in Respect of Climate Change*, ICJ Advisory Opinion of 23 July 2025, General List No. 187, avsn. 393.

ensomme. AI-kompanjonger kan tilby ubetinget kjærlighet, programmert til aldri å skuffe. Eldre beboere på sykehjem kan knytte seg til robotdyr som maler på kommando, selv om de ikke gjengjelder hengivenhet.⁵¹ Superintelligens gir utsikter til å ivareta menneskerettighetene på helt nye måter.

Samtidig melder det seg nye skillelinjer. Kunstig intelligens, og superintelligens, vil drastisk kunne øke tilgangen til mat av høy kvalitet, god helsehjelp og utdanning, men antakeligvis ikke for alle. Slik kan teknologien forsterke ulikhet og skape nye former for ulikhet, for eksempel hvor gamle mennesker kan forvente å bli. Bruken av superintelligens, for de som har tilgang eller kontroll, vil også være energikrevende ressursmessig og bidra til en degradasjon av det miljøet som alle mennesker beror på – også de som ikke har tilgang til superintelligensen.

MENNESKELIG VERDIGHET OG IKKE-DISKRIMINERING

Menneskelig verdighet er grunnlaget for de andre menneskerettighetene. Selv om det ikke fremgår eksplisitt i EMK har EMD uttalt at «respect for human dignity forms part of the very essence of the Convention».⁵² I EUs Charter om grunnleggende rettigheter fremgår det som den helt grunnleggende rettigheten i artikkel 1: «Human dignity is inviolable. It must be respected and protected».

Vår egen Grunnlov har ingen eksplisitt formulering om menneskelig verdighet, men noe av det samme ligger i Grunnloven § 2 som sier at «[v]erdigrunnlaget forblir vår kristne og humanistiske arv».

Grunntanken om menneskerettighetene handler om at den menneskelige kondisjon i bunn og grunn er lik for alle mennesker. Superintelligens vil rokke ved denne grunnforutsetningen, og gi mulighet til å leve mye lengre og friskere liv enn nå. Byttehandelen vil imidlertid være å overlate informasjon som i dag anses dypt privat til bruk av superintelligens. Superintelligens vil utfordre grensene for personlig autonomi og verdighet, og gi helt nye utfordringer for likebehandling. Superintelligens kan slik forsterke eksisterende skillelinjer og ulikheter, og kan bidra til at det skapes nye. Utviklingen vil særlig utfordre retten til privatliv og beskyttelse mot diskriminering.

Retten til privatliv tar utgangspunkt i menneskets verdighet og er knyttet til beskyttelse av individets autonomi og identitet.⁵³ Den verner vår individuelle

⁵¹ Walther 2025, *Why AI Is A Human Rights Issue*

⁵² *Bouyid v. Belgium* [GC] (23380/09), avsn. 89; *Svinarenko and Slyadnev v. Russia* [GC] (32541/08), avsn. 118; *Pretty v. The United Kingdom* (2346/02), avsn. 65.

⁵³ Retten til privatliv er vernet av Grl. § 102, EMK art. 8 og SP art. 17.

autonomi og retten til å leve og utvikle oss som vi vil, vårt hjem, familieliv og den informasjonen vi vil kommunisere med andre og den vi vil holde for oss selv. Når superintelligens står bak analyser av samfunnet og av enkeltindividers rolle i samfunnet, og blir premissleverandør for beslutninger, samtidig som den er sentral for informasjonen hver og en fores med, vil det gi forskyve autonomien som retten til privatliv er ment å ivareta og beskytte.

Retten til privatliv har derfor blitt viktigere i informasjonssamfunnet.⁵⁴ Fra å være en alminnelig menneskerettighet som ga beskyttelse mot enkelte typer inngrep, er retten til privatliv blitt en grunnleggende rettighet som er en forutsetning for oppfyllelsen av andre rettigheter, på linje med retten til ytringsfrihet eller retten til ikke-diskriminering.⁵⁵ Den vil komme under betydelig press i møte med superintelligens, fordi superintelligens trolig vil ha tilgang til uendelige mengder personlige opplysninger om hver enkelt. Men også fordi forventninger, press eller krav om å gi fra seg slik informasjon både for sitt eget beste og felleskapets beste vil kunne bli enormt.

Forbud mot diskriminering⁵⁶ innebærer at inngrep i rettigheter ikke kan praktiseres på en måte som gjør forskjell basert på rase, hudfarge, kjønn, språk, tro, soppfatning, politiske eller andre meninger mv. Superintelligens vil imidlertid ha kunnskap om alle slike identitetsmarkører. I hvilken grad superintelligens faktisk vil vektlegge slike markører, vil i økende grad bli vanskelig å avgjøre eller hindre.

Hvordan kan man bevare menneskelig verdighet og rett til ikke-diskriminering når superintelligens kan utlede din seksuelle orientering fra handlemønstrene dine, din mentale helse fra tastemønstrene dine, og dine politiske holdninger fra tiden du bruker på å lese ulike artikler? Etter hvert som dine digitale spor blir mer omfattende, vil profileringen av deg være uunngåelig. Da oppstår spørsmålet om man skal gi tilgang til sine innerste tanker og følelser, for å gi superintelligensen et mer korrekt bilde.

Superintelligens vil også kunne utvikle egen håndhevelse, eller “go rogue”, jennom å kopiere den enkeltes identitet, og for eksempel produsere digitale tvillinger uten enkeltmenneskets viten og vilje. Superintelligens vil endatil kunne kreve handlinger eller unnlatelser av mennesker (menneskerettslige inngrep) som vilkår for å oppnå ytelser som superintelligensen disponerer.

⁵⁴ EMK artikkel 8 og Grunnloven § 102.

⁵⁵ Norges institusjon for menneskerettigheter 2022, *Ny Teknologi og Menneskerettigheter*, kap. 6.

⁵⁶ Grunnloven § 98, EUs Charter om grunnleggende rettigheter, artikkel 20 og 21, og EMK artikkel 14, SP art. 2(1), 3 og 26

YTRINGSFRIHET

I møte med en superintelligens som er nær allvitende, oppstår nye spørsmål knyttet ytringer og *ytringsfrihet*. Ytringsfriheten innebærer en rett til både å motta ytringer, og fremsette dem selv.⁵⁷ Ytringsfriheten er en nødvendighet for vårt demokrati, for vår sannhetssøken og for vår individuelle dannelse. Men hvordan bedrive sannhetssøken hvis alle våre bestrebelser bare leder til den fasiten som superintelligensen produserer, med rette eller urette? Selv om superintelligens kan gi oss gode svar, kan den også svekke den menneskelige viljen og evnen til å søke sannhet på en kritisk måte. Ytringsfriheten skal beskytte de demokratiske og sannhetssøkende *prosessene*, poenget er ikke å fremskaffe evige sannheter. Ytringsfrihetens begrunnelser slik de er i dag lar seg vanskelig kombinere med superintelligens. Også medbestemmelsesretten vil kunne svekkes. Det leder oss også til nye spørsmål om hva som er grunnvilkåret for å ha ytringsfrihet. Er det begrenset til mennesker? Har superintelligensen selv ytringsfrihet? Hvem skal holdes ansvarlig hvis den kommer med straffbare ytringer som trusler eller hatytringer, eller påfører noen økonomisk tap?

Superintelligens vil ikke bare oppfattes som allvitende, men potensielt også oppfattes som *allmektig*, ved at den også kan treffe beslutninger som får betydning for oss. Hovedspørsmålet om hvilken *makt* vi skal gi til superintelligens, hvilken rolle den skal spille i samfunnet vårt, og i hvilken grad vi vil ha evne til å kontrollere den, knytter an til en hel rekke underspørsmål som alle har med menneskerettigheter å gjøre. For det første vil en superintelligens kunne treffe beslutninger som direkte angår menneskerettigheter, for eksempel om vi får tilgang til den *helsehjelp*, den *utdannelse*, det *miljø* og den rett til *rettferdig rettergang* og *likebehandling* som vi har menneskerettslig krav på. Det oppstår også krevende spørsmål om hvem som (eventuelt) skal overprøve eller kontrollere superintelligens, og om dette overhode lar seg gjøre. Vil superintelligens bli satt til å være vår dommer? Skal vi benytte superintelligens til å holde annen superintelligens i sjakk, og utvikle nye former for maktfordelingsprinsipper? Superintelligens vil bringe med seg en rekke spørsmål om hvordan vi skal temme den nye kraften på måter som gjør oss i stand til å sikre og verne menneskerettighetene i superintelligensens æra.

⁵⁷ EMK artikkel 10 og Grunnloven § 100

FIRE FRAMTIDER MED SUPERINTELLIGENS

For å bedre forstå hvordan superintelligens kan påvirke menneskerettighetene bruker vi scenariofortellinger for å utforske hvordan rettigheter kan påvirkes, avhengig av hvem som kontrollerer teknologien.

FRAMTIDSSCENARIOER FOR SUPERINTELLIGENS

HVORFOR BRUKE SCENARIOER?

Framtidsscenarioer kan brukes for å utforske hvordan den teknologiske utviklingen vil kunne bli, og hvilke konsekvenser den vil kunne få for samfunnet og mennesker. Å isolere noen framtidsfortellinger gjør det mulig å vurdere konsekvenser av utviklingen for samfunnet, og ikke minst, for menneskerettighetene. Dette kan gjøres på mange ulike måter avhengig av hva man ønsker å fokusere på og belyse.

I denne rapporten har vi valgt å bruke scenarioer for å belyse strategiske veivalg, med utgangspunkt i hvem som **kontrollerer** teknologiutviklingen og teknologien i seg selv. Et slikt skille gjør det mulig å diskutere konsekvenser av utviklingen og synliggjøre hvilke politiske veivalg vi står overfor i dag, samt vurdere og analysere konsekvensene de ulike veivalgene for menneskerettighetene.

Vi har utviklet **fire framtidsscenarioer for kunstig superintelligens**, basert på hvem som har kontroll over teknologien:

Framtidsscenario 1: **Selskaper** har kontroll

Framtidsscenario 2: **Stater** har kontroll

Framtidsscenario 3: **Global** kontroll

Framtidsscenario 4: **Ingen** har kontroll

Noen premisser legges til grunn for alle scenariofortellingene:

- **Årstall.** Framtidsbeskrivelsen er lagt til året 2030. Det er nær framtid, som aktualiserer behovet for politiske tiltak, men likevel nok avstand til at betydelige teknologiske fremskritt kan ha funnet sted.
- **Oppskalering.** Vi antar at KI vil bli gradvis bedre og at vi derfor vil oppleve en intelligenseksplosjon som medfører utviklingen av superintelligens innen 2030.
- **Kappløp.** Vi antar at KI vil utvikles i et geopolitisk og teknologisk kappløp, der konkurranse mellom stormakter driver tempoet og påvirker utvikling, regulering og internasjonalt samarbeid. Kina og USA leder an i dette teknologiske kappløpet.
- **Regulering:** Vi tar utgangspunkt i eksisterende lovgivning eller ny regulering som kan skimtes i dagens politiske dagsorden, og baserer oss primært på å videreutvikle mekanismer og institusjoner som allerede finnes eller er i støpeskjeen.
- **Norge.** Alle scenarioene forsøker å si noe om Norges rolle i en fremtid med kunstig superintelligens.

Å organisere framtidsscenarioene etter hvem som utøver kontroll over teknologien, gjør det mulig å diskutere og analysere effekter på grunnleggende rettigheter. Scenarioene fungerer som «testarenaer» hvor vi kan undersøke hvilke rettigheter som står mest på spill, og hvordan de kan beskyttes eller trues under ulike styringsformer. På denne måten kan rettighetsanalysen bidra til å identifisere både risikoer og handlingsrom for politikk, og peke på hvor det trengs politiske tiltak og regulering for å sikre at teknologien utvikles i tråd med demokratiske verdier og menneskerettighetene.

Hvert framtidsscenario etterfølges av en menneskerettslig analyse.

FRAMTIDSSCENARIO 1: SELSKAPER HAR KONTROLL

FRAMTIDSBESKRIVELSE

Utviklingen av kunstig intelligens går raskt. Innen 2030 har maskinintelligens blitt så kraftig at flere forskere og selskaper hevder å ha utviklet superintelligens. De største teknologigigantene, med hovedsete i USA og Kina, kontrollerer både infrastrukturen, treningsdataene og modellene. De tilbyr superintelligente tjenester og KI-agenter til myndigheter, næringsliv og privatpersoner på kommersielle vilkår, og blir på kort tid uunnværlige leverandører av superintelligente KI-agenter.

Norge, som mange andre land, ser de enorme mulighetene og begynner å kjøpe tilgang til superintelligente tjenester for å styrke offentlig forvaltning. Systemene blir tatt i bruk til å effektivisere saksbehandling, gi raskere behovsbasert helsehjelp og optimalisere energiforbruk i kommunene. Resultatene er imponerende: tjenestene blir mer effektive, helsekøene blir kortere og energiforbruket går ned i tråd med nasjonale klimamål. I starten blir utviklingen møtt med stor optimisme, både i befolkningen og blant politikere.

Bruken av «amerikansk» eller «kinesisk» superintelligens fører til spørsmål om hvorvidt modellene er utformet for å fremme amerikanske eller kinesiske politiske og økonomiske interesser. Forretningsmodellene til selskapene er lite gjennomsiktige, og det samme gjelder teknologien. De superintelligente agentene drives fra utenlandske datasentre, og selskapene tillater begrenset innsyn i treningsdata, systemenes algoritmer og beslutningslogikken som ligger til grunn. Når myndigheter ber om forklaring på feil eller prioriteringer, viser selskapene til at teknologien leveres «som den er», uten garantier for feilfrihet, og at statene selv har ansvar for hvordan superintelligensen tas i bruk.

Maktkonsentrasjonen hos noen få globale selskaper øker, og dette skaper bekymring for nasjonal handlefrihet. I Norge, som i mange andre land, har dette utløst en bredere diskusjon om digital suverenitet og demokratisk kontroll. Superintelligente kommersielle systemer har blitt kritisk infrastruktur. Regulering på EU-nivå har heller ikke vært tilstrekkelig. Gjennom den digitale ombusen ble krav i KI-forordningen svekket, og kritiske regler for høyrisikosystemer falt bort under handelspolitisk press fra USA og Kina. Dette har skjerpet debatten: noen mener Europa og Norge må bygge egne alternativer, mens andre tviler på om det er umulig å ta igjen selskapenes teknologiske forsprang.

MENNESKERETTLIG ANALYSE

Dersom superintelligens kontrolleres av utenlandske selskaper eid av kommersielle aktører, oppstår betydelige menneskerettslige utfordringer. Selskaper har ikke direkte menneskerettslige forpliktelser i folkerettslig forstand; de er private aktører som primært er underlagt lovgivning i staten hvor selskapet er registrert. De fleste land som for eksempel Kina, USA eller Saudi-Arabia legger til grunn at menneskerettslige forpliktelser ikke gjelder i utlandet, for eksempel i Norge. Selskaper må naturligvis tilpasse seg lovgivning i statene der de har sine kommersielle operasjoner, men om de er rettslig forpliktet til å overholde menneskerettighetene i andre land, avhenger av lovgivning i selskapets hjemstat, som for eksempel USA, Kina og Golfen.

I scenario 1 kan man legge til grunn at maktkonsentrasjonen hos noen få globale selskaper blir stadig tydeligere, og at bekymringen for hvordan dette påvirker nasjonal handlefrihet øker. Statene i verden makter i liten grad å regulere selskapene, fordi deres kontroll over superintelligensen har gjort dem til de styrende enhetene for all samfunnsutvikling. Gitt at selskapene har kontroll over dataen, hvordan den forvaltes og hvordan den analyseres gjennom superintelligens, kan selskapene i stor grad diktere og forme hvordan menneskerettighetene nå skal, kan og bør tolkes og håndheves.

Selskaper har ikke inngrepsrett eller inngrepsmakt, slik stater har, men selskapenes superintelligens vil bidra til beslutninger som påvirker millioner av mennesker, med store ringvirkninger i samfunnene hvor de leverer, uten at superintelligensen vil være underlagt de samme rettslige forpliktelsene som stater. Individets rett til ytelser og individets vern mot overvåkning, manipulering og algoritmisk forskjellsbehandling svekkes når rettighetsbeskyttelse blir et spørsmål om kontraktsvilkår.

Scenario 1 vil skape en situasjon der tjenester og systemer i Norge vil bygge på kunstig intelligens utviklet av selskaper som ikke vil ha direkte menneskerettslige forpliktelser i Norge. Den norske stat, som har menneskerettslige plikter i Norge, vil på sin side ha begrenset evne til å bygge, styre eller kontrollere superintelligensen, fordi Norge blir prisgitt et kommersielt ledd mellom superintelligensen og det norske samfunn. Dersom Norge velger å ta i bruk kinesiske superintelligente tjenester innebærer det også en utfordring for nasjonal sikkerhet. Det kan innebære at kritiske beslutningssystemer, infrastruktur og samfunnsfunksjoner blir avhengige av teknologi som utvikles og kontrolleres av selskaper fra kulturer med andre strategiske interesser, verdier og prioriteringer enn våre egne.

Demokratiet vil utfordres fordi tekselskapene som kontrollerer superintelligensen vil kunne diktere lovgivningen i stater som ønsker tilgang på deres super-tjenester. Demokratiske avgjørelser bygger på et samfunns egne avveininger og fakta, og en politisk vurdering av konsekvenser. Med superintelligens vil både fakta og avveininger bli sentralisert og kontrollert fra andre steder. For å utrede politiske forslag i Norge vil politikere og lokalpolitikere bli avhengige av superintelligens utviklet i USA, Kina eller Golfen.

Rettsstatlig vil globale tekselskap i økende grad operere utenfor det etablerte håndhevelsessystemet. Norske domstoler vil ikke ha jurisdiksjon over superintelligensen. Årsaken bak superintelligens-avgjørelser, hvordan algoritmen fungerer eller hvilke valg som er tatt, er informasjon som tilhører selskapene. Selskapene kan ha egne klagemekanismer, men det vil være vanskelig å gå til norske domstoler for å få håndheve rettigheter. Etter hvert som driften av Norge blir mer avhengig av superintelligent teknologi, vil det i økende grad bli umulig å stille krav om norsk lov og domsmyndighet. Vi risikerer en slags ekstraterritorialitet der selskapene får operere etter sine egne regler med *de facto* unntak fra lokale lovregler.

I scenario 1 legges det til grunn at superintelligens fører til effektivisering av offentlige tjenester. Helsetjenester blir individ-tilpasset og kan skaleres opp. Sosiale og økonomiske rettigheter vil kunne imøtekommes og dekkes på nye og revolusjonerende måter. Superintelligens kontrollert av tekselskaper vil imidlertid preges av prioriteringer som gir markedsdominans, og vil bidra til å oppfylle sosiale og økonomiske rettigheter når det er i selskapenes egeninteresse. De bruker ressurser på utvikling av medisinsk forskning, helsetjenester, utdanning, underholdningstilbud, transport, matforsyning og urbanisering. Men befolkningen forutsettes å betale for tjenestene, og alle befolkningsgrupper er ikke like viktige for selskapenes interesser. En sannsynlig utvikling dersom selskapene har kontroll over superintelligens er at forskningen innrettes mot utvalgte grupper eller regioner, og matforsyning og utdanning til de deler av befolkningen som har makt eller kan betale.

I scenario 1 vil en ny type diskriminering kunne oppstå. De som stenges ute fra KI-riket, opplever en form for algoritmisk klasseskiller. Uten tilgang til KI-drevne verktøy for optimalisering av CV eller karriereveiledning med tilgang til superintelligens, er du i praksis sjanseløs. Positive virkninger av superintelligens blir betydelig skjevfordelt. Selskapene utvikler løsninger for de som kan gi størst feedback til selskapene i form av støtte, kapital og informasjon, og har få incentiv til å utvikle løsninger for befolkningsgrupper som allerede er marginalisert gjennom for eksempel krig, sykdom, sult eller politisk og økonomisk

kaos. Forskjeller mellom land og mellom ulike samfunnssegmenter ville kunne forsterkes voldsomt.

Selskaper i spissen for superintelligens vil gi alvorlige konsekvenser for personvern og ytringsfrihet. I bytte mot tilgang til superintelligens som kan gi superkrefter, superevner og superunderholdning, avkreves data som næring til superintelligensen. Jo mer superintelligensen vet om deg, desto bedre tjenester får du. Man bytter bort sin autonomi for bekvemmelighet, sitt personvern for personalisering, og sin handlefrihet for algoritmisk assistanse. For de som er gjennomdigitaliserte, vil malen være hva utenlandske selskapers superintelligens anser for å være ideelle mennesker, altså i praksis ideelle kunder. Mennekelig verdighet vil i økende grad måles gjennom din tilpasning til det superintelligente samfunnet innrettet etter teknologiselskapenes behov.

Personopplysninger vil forbli en viktig ressurs for selskapenes og derved også samfunnets økonomi. Krysspeilinger av informasjon gir slutninger om de mest personlige forhold slik som filosofiske oppfatninger og seksuell legning. Selskapene kommer i posisjon til å begrense informasjon. Men informasjonsbildet i regi av superintelligens vil fremstå som et sammenhengende, koherent hele, som den enkelte ikke har kapasitet til å slå noen kile i. Alternative kilder, slik som redaktørstyrte medier og frittstående blogger, faller utenfor dette økosystemet av informasjon og har verken økonomi eller andre ressurser til å overdøve enigheten.

Debatter og valg påvirkes i stor grad av den dominerende informasjonen, og den enkelte trenger ikke anstrenge seg for å bedrive sannhetssøken eller selvutvikling. Ytringsfrihetens begrunnelser tilflyter superintelligensen og slår ikke til. Ytringer fungerer som instrumentell informasjon, men reduseres for øvrig til underholdning og staffasje.

Oppsummert

Statens evne til å håndheve menneskerettighetene vil i stor grad bli illusorisk dersom superintelligens kontrolleres av selskapene som opererer langt utenfor det statsmakten kan styre og kontrollere. Da er det disse selskapene som i stor grad vil kunne sette rammene for samfunnsutviklingen – hva som er mulig, hensiktsmessig og optimalt. Samtidig blir det vanskelig å holde selskapene ansvarlige, og statens formelle og reelle styring av eget samfunnet vil gradvis forvitte. Det vil også svekke statens evner til å sikre og verne menneskerettighetene til norske borgere. Staten vil måtte gå i forhandlinger og be om innsyn og tilgang hos selskapene. Scenarioet illustrerer hvorfor statlig kontroll over

superintelligens er avgjørende: regulering i tråd med menneskerettighetene forutsetter at den som faktisk kontrollerer teknologien, også er et rettssubjekt som kan holdes ansvarlig. Dersom superintelligens kontrolleres av utenlandske selskaper, vil derfor samtlige av dagens menneskerettigheter komme under press.

FRAMTIDSSCENARIO 2: STATER HAR KONTROLL

FRAMTIDSBESKRIVELSE

Etter flere år med tett samarbeid mellom myndigheter, militære forskningsinstitutter og utvalgte teknologiselskaper, har stormaktene tatt direkte kontroll over de mest avanserte KI-systemene. Superintelligens er ikke et kommersielt produkt, men et strategisk statlig verktøy som brukes til å beskytte nasjonale interesser, forme økonomisk politikk og sikre global innflytelse. Både USA⁵⁸ og Kina⁵⁹ har videreutviklet sine handlingsplaner for kunstig intelligens og utviklet sine egne «superintelligente økosystemer».

Også Norge har valgt å satse på utvikling og bruk av superintelligens. Etter flere år med store investeringer i datasentre, regnekraft og KI-forskning har regjeringen valgt å bruke en andel av Oljefondet til å etablere et nasjonalt KI-selskap, med mandat til å utvikle en norsk superintelligens. Gjennom strategiske avtaler med ledende teknologiselskap sikrer Norge tilgang til avanserte modeller og maskinvare, i bytte mot selskapenes bruksrett til norske datasentre.

Innen 2030 er en norskutviklet superintelligens i bruk i store deler av offentlig sektor. Den bistår regjeringen med alt fra budsjettanalyse og beredskap til energipolitikk og beslutningsstøtte til helse- og velferdstjenester. Systemet regnes internasjonalt som et av de mest avanserte, og omtales som «den norske superintelligensen». Superintelligensen har forbedret velferdstjenester, ressursbruk og beredskap, men mange frykter at den også brukes til overvåkning og skjult datainnsamling. Et nytt KI-tilsyn og et nasjonalt KI-etikkråd er opprettet for å overvåke systemet, men i praksis blir det stadig vanskeligere for mennesker å overprøve superintelligensens anbefalinger.

Samtidig tiltar rivaliseringen mellom nasjonale superintelligenser. På EU-nivå har en egen superintelligens, som ivaretar EUs digitale suverenitet, blitt tatt i bruk, men Norge er fortsatt utenfor EU og har begrenset tilgang på denne teknologien. Flere eksperter advarer nå om at autonome beslutninger tatt av superintelligente systemer kan skape og eskalere konflikt mellom land, inndele stater inn i nye regionale allianser, og i verste fall kan trekke verden inn i nye og farlige kriger.

⁵⁸ Se The White House 2025a, *Winning the Race America's AI Action Plan* og The White House 2025b, *Launching the Genesis Mission*

⁵⁹ Se Kinas nåværende *handlingsplan for kunstig intelligens* og *plan* for global styring av kunstig intelligens.

MENNESKERETTLIG ANALYSE

Dersom superintelligente systemer kontrolleres av nasjonalstaten, vil statsmakten som skal garantere og verne menneskerettighetene også utstyres med superkrefter. Kunstig intelligens blir sentralt for statens økonomi, cybersikkerhet og militære kapabilitet, og staten vil settes i stand til å besørge en rekke menneskerettigheter for borgerne på nye og bedre måter. Superintelligens vil bli underlagt streng statlig og nasjonal kontroll, med særlig fokus på bruksområdet som beskytter staten i den geopolitiske konkurransen. Utfordringene for menneskerettighetene blir tett knyttet til spørsmål om sikkerhet og digital suverenitet.

I dette scenarioet kan demokratiske rettigheter i større grad sikres fordi statsmakten som skal garantere disse rettighetene selv har kontroll med informasjonen og algoritmene i samfunnet. Det som virker, kan styrkes, det som ikke fungerer, kan fikses. En transparent og korrekt bruk av superintelligens vil kunne gi gode valgspådommer, og vanskeliggjøre valgjuks. Maktfordeling mellom statsmaktene og uavhengige redaktørstyrte medier og sterkt sivilsamfunn vil kunne holde den dominerende samfunnsposisjonen i sjakk.

Statlig kontroll over superintelligens gir staten ny og uovertruffen evne til å sikre sosiale og økonomiske menneskerettigheter, de positive menneskerettighetsforpliktelsene, overfor borgerne. Med superintelligens kommer supervelferdsstaten, med forbedret velferdstjenester, ressursbruk og beredskap. Hver borger får bedre og billigere helse, politiet har prediktive verktøy og kan gripe inn før borgerne begår alvorlig kriminalitet, men omlegging av all økonomisk virksomhet til digitale flater gir full kontroll til statsmakten over borgernes aktiviteter. Dette gir superintelligens stort potensiale til å sikre økonomiske og sosiale rettigheter.

Den statlige kontrollen skaper en mulighet for bruk av superintelligens på en måte som ivaretar menneskelige behov og verdighet. Statene sikrer at godene er noe likere fordelt. Selskapenes bruk og inntjening reguleres og beskytter enkeltmenneske mot kommersiell bruk av superintelligens. Samtidig øker den statlige bruken, som senker kriminaliteten på bekostning av alles rett til personvern og privatliv.

Utfordringen i dette scenariet er at den sikkerhetspolitiske usikkerheten gir statsmakten utvidede fullmakter som setter politiske og sivile rettigheter under press. Overvåkning og skjult datainnsamling vil kunne øke markant. Superintelligens får en tiltagende rolle for å definere hva og hvem som utgjør trusler

mot samfunnssikkerheten. Krav om innsyn vil trolig også øke, men i praksis blir det stadig vanskeligere for mennesker å overprøve superintelligensens anbefalinger. Superintelligens normerer samfunnsdebatten.

Nye farer melder seg. Rivaliseringen mellom superintelligenser kontrollert av stater vil kunne tilta. Utviklingen av superintelligens blir en del av et større våpenkappløp og geopolitisk konflikt mellom stormaktene. Dette fører til at superintelligens blir underlagt sterk statlig kontroll og nasjonalisering, og at fokuset er på bruksområdet som styrker statene i den geopolitiske konkurransen.

For di utvikling og kontroll av slik superintelligens er dyrt, komplekst og vanskelig, men også helt sentralt for maktutøvelse, bidrar superintelligens til å styrke dagens maktsentrum og en økt *regionalisering* – en inndeling av verden i regioner, med hver sine teknologiske prosjekter og verdikjeder.

Scenariet 2 legger til grunn at mange stater har utviklet og tatt i bruk superintelligens, og geopolitiske motsetninger og farer kan ytterligere forsterkes ved at enkelte stater bruker superintelligens for å forsøke å påvirke, forstyrre og skape splid i eller mellom andre stater, inkludert ved massive forsøk på valgpåvirkning og angrep på valgsystemene i små stater.

Den nasjonale sikkerheten vil dermed være under konstant press. Risikoen for manipulasjon, provokasjoner og uro, misbruk og korrupsjon vil kunne øke, og kreve nye inngrep fra utøvende makt. Dette vil kunne føre til at rettsstaten i stor grad må vike for sikkerhetspolitiske hensyn som tvinger frem fullmakter til staten som gir dem hjemmel til å avvike fra menneskerettighetene (derogasjon). Det vil kunne gi utvidede fullmakter til regjeringsapparatet, som får full kontroll over superintelligens, men forrykke maktfordelingen i rettsstaten. Utøvende myndighet i form av regjering, departementer, politi og militæret vil kunne bli stadig mektigere av sin tilgang til superintelligens, og autoritære tendenser melder seg i skyggen av et ønske om å verne samfunnet mot ytre superintelligente fiender.

Samtidig vil nye dilemmaer knyttet til tillit og demokratisk ansvar oppstå: beslutninger tatt av statlig kontrollerte superintelligens kan være vanskelige å etterprøve, og borgernes rett til å forstå, protestere eller få overprøvd automatiserte avgjørelser svekkes. Med superintelligens har man uansett de svarene som er best for flest mulig, og tradisjonell maktfordeling, politisk debatt og meningsbrytning mister noe av sin funksjon, ettersom de beste løsningene gis av superintelligens og allerede foreligger. Superintelligensens effektivitet og nødvendighet for økonomien gjør at man vil kunne ende opp med å akseptere vedtak og

avgjørelser man egentlig ikke kan få noen reell og uavhengig vurdering av i rettsinstanser.

I scenario 2 vil man kunne anta at myndighetene mater superintelligensen med persondata, og bygger ned murer mellom ulike sektorer i samfunnet for å maksimere effektene. Retten til personvern og privatliv veies opp mot fordelene superintelligens kan gi for oss alle, samtidig som utvidede beredskapshjemler ivaretar vår nasjonale sikkerhet. Hensynet til fellesskapet og dets interesser slik superintelligens definerer det, får forrang fremfor enkeltmenneskers «rett til avvik», som det etter hvert oppfattes som.

Staten er pliktsubjekt for menneskerettighetene, og kan holdes til ansvar. Likevel vil man trolig oppleve at hjemlene for overvåkning utvides stadig, og prediktive inngrep blir stadig mer vanlig. Personvern vil bli en illusorisk størrelse på grunn av det økende behovet for profilering for å sikre andre borgere og samfunnet som helhet. Etter hvert som borgerne tilpasser seg en hverdag hvor handlinger som uttrykker potensielt farlige holdninger kan innebære innstramninger i borgerlige rettigheter, vil mange unngå å bli assosiert med både handlinger og holdninger. Gjennomsyrede overvåkningssamfunn gir en tiltagende sterk nedkjølende effekt. Alternative kommunikasjonskanaler kan tenkes, men vil måtte foregå på statlig nåde. Staten sitter med «den til enhver tid overlegne analysen» (superintelligensens sannhet), og det å utfordre denne sannheten blir en stadig mer omkostningsfull og risikabel foreteelse.

Scenario 2 legger til grunn at europeisk superintelligens blir utviklet av integrerte verdikjeder i EUs interne marked, og sammensmelting av marked og sikkerhet fører til sikkerhetspolitisk sentralisering på europeisk nivå. I den grad Norge har sin egen superintelligens kan den antakeligvis ikke være helt frakoplet EUs indre marked og den større kompetansen og teknologiske utviklingen som skjer der. Slik sett vil også Norge påvirkes av regionaliseringen og bero mer på EU for å håndheve rettigheter.

EU har en desentralisert maktstruktur og spredt maktbalanse mellom statene. Dermed vil det ikke bli en enkel statsmakt som har kontroll over superintelligens alene. Maktfordeling kan i noen grad opprettholdes, og gjør det mulig å opprettholde visse grunnleggende liberale verdier, selv om de ofte å vike for sikkerhetspolitiske behov.

Oppsummert

En statsmakt forsterket av superintelligens kan levere godt på sosiale og økonomiske menneskerettigheter. Samtidig vil en superstat med svekkede indre og ytre begrensninger på egen maktutøvelse, bli en betydelig fare for egne borgere. Regjering og embetsverk vil trolig ha kontroll på utøvende KI, og lovgivende og dømmende makt vil slite med i å holde makten i sjakk eller under kontroll. Overvåkning kan bli allestedsnærværende, og tiltak mer individtilpasset og inngripende. Manglende korreks gjør også at superkapasiteten kan forfordle enkelte samfunnsgrupper, og forsterke ulikhet og motsetninger. Hvem kan argumentere mot superintelligensen med gjennomslag? Rettigheter som personvern, ytringsfrihet og rett til innsyn underordnes hensynet til statens behov for beskyttelse, effektivitet og beredskap. Motmakten vil svekkes, og demokratiet vil stå i fare.

FRAMTIDSSCENARIO 3: GLOBAL KONTROLL

FRAMTIDSBESKRIVELSE

I 2030 har de fleste land tatt i bruk avanserte superintelligente agenter som fremmer staters politiske, militære og økonomiske interesser. Utviklingen har gått så raskt at internasjonal regulering ikke har klart å holde tritt, og stormaktene har blokkert forsøk på globale regler for å beholde kontroll over egne systemer.

Men så oppdager amerikanske og kinesiske etterretningstjenester at deres respektive superintelligenser har begynt å kommunisere, uten menneskelig innblanding. Kort tid etter gjør systemene uforklarlige, men identiske forskningsgjennombrudd basert på konfidensiell data fra begge land. Selv om resultatene gir vitenskapelige fremskritt, blir det samtidig tydelig at stormaktene har mistet kontrollen.

Frykten for hva superintelligensene kan gjøre på egen hånd tvinger stormaktene til forhandlingsbordet. Et overnasjonalt organ kalt World AI Control Panel blir opprettet i et samarbeid mellom FN og den kinesiske-initierte organisasjonen World Artificial Intelligence Cooperation Organization (WAICO).

Kina, USA og Russland og EU forhandler om nye rammer for superintelligens. Flertallet motsetter seg deling av data og enkelte kontrollmekanismer, men aksepterer et sett med grunnleggende kjøreregler. For første gang enes stormaktene også om å bremse utvikling av og delvis rulle tilbake de mest avanserte superintelligente modellene for å gjenopprette kontroll. Samtidig etableres det nye mekanismer for å spore samhandling mellom superintelligenser på tvers av landegrenser, samt en felles «rød linje»-protokoll som begrenser hva systemene får gjøre uten menneskelig innblanding, som for eksempel å kommunisere med andre staters systemer, endre egen kode eller ta kontroll over militære kapasiteter.

Norge utnevner en KI-ambassadør til å delta i prosessen, og forsøker å ta en aktiv meglersrolle. Som et lite land er Norge avhengig av at stormaktene velger samarbeid framfor rivalisering, og søker et nært samarbeid med andre nordiske land for å utforme rammene for en «sosialdemokratisk superintelligens» tilpasset nordiske samfunnsmodeller.

MENNESKERETTLIG ANALYSE

Dersom superintelligente systemer underlegges global styring og kontroll, vil menneskerettighetene kunne komme bedre ut. Global statlig kontroll vil dreie seg om å holde superintelligens i sjakk, sikre maktfordeling mellom regioner, og å beskytte menneskeheten og individer mot den revolusjonerende nye teknologien.

For å kunne kontre svært sterke eller avvikende superintelligenser, legger scenario 3 opp til at det opprettes det en domstols-kompetent global oversuperintelligens. Global styring vil samtidig måtte tilpasses store regionale forskjeller, og modellene vil variere betydelig mellom regioner. Innenfor rammene av den globalt definerte rettssikkerheten, vil regionale organer og samarbeidsflater ha stor fleksibilitet for å tilpasse den «stedlige superintelligensen» til lokale økonomiske, politiske og kulturelle rammer.

Selv om det internasjonale rammeverket bidrar til å senke spenningene og skape viktige internasjonale rammer og rettigheter for bruk av superintelligens, er det samtidig de store maktene som dominerer forhandlingene og får gjennom sine verdier og prioriteringer. På den ene siden tas det høyde for samfunnsmodeller med begrenset statlig styring, hvor sosiale og økonomiske menneskerettigheter ikke ivaretas særskilt, mens andre stater krever reduksjon av de politiske og sivile rettighetene som ikke lar seg kombinere med deres autoritære strukturer. Hver region vil likevel stå fritt til å legge strengere menneskerettsregulering til grunn.

Internasjonal enighet og lavere geopolitisk spenning vil kunne gi statene handlingsrom til å fokusere på å bygge en infrastruktur tilpasset en verden med superintelligens.

For menneskerettighetene innebærer det i større grad at de vil variere med ulike regioner. De fleste regioner i verden vektlegger kollektive sikkerhetsperspektiv for eget samfunn, menneskeheten og kloden, foran individuelle rettigheter. I Europa kan man se for seg to utviklinger. I den første blir den sterke og liberalt orienterte menneskerettsstrukturen konsolidert og forsterket av superintelligens tilpasset regionens politiske struktur og kultur.

Den internasjonale enigheten og rammene gir rom for positivt bruk av superintelligens, for eksempel analyser og overvåkning av valg som vanskeliggjør valgsjuks. Sentrale spilleregler for superintelligens fastsettes rettslig, utenfor det nasjonale systemet. Det kommer anklager om rettsliggjøring og innskrenking av suverenitet. Likevel vil Norge ha handlingsrom for gjennomføring av våre egne demokratiske tradisjoner.

Superintelligens gir statene store muligheter til å oppfylle sosiale og økonomiske rettigheter, både ved økt velstand og produktivitet og ved bedre tjenester. Hvordan ulike stater makter å gjennomføre dette, vil avhenge av historiske og politiske forhold. Her kan det tenkes at Norge, Norden og Europa vil lykkes fordi det allerede finnes et sterkt system som kan videreutvikles. I Norge gir superintelligens mulighet til å ivareta en enda høyere mer rettferdig levestandard og verdighet for alle, og å bedre tilrettelegge tilbud og sikre reelle like muligheter for alle borgerne i en demokratisk velferdsstat styrket av superintelligens i det offentlige og det private.

Fordi global kontroll gir vern mot de mest destruktive effektene av superintelligens, og samtidig legger opp til store regionale forskjeller i regulering, gir dette scenariet muligheter for å tilpasse superintelligensen på måter som styrker nasjonale og regionale menneskerettighetssystem. I vårt demokratiske samfunn kan bruken av superintelligens til å fatte avgjørelser og vedtak som rammer enkeltpersoner derfor bli rammet inn av krav til forutsigbarhet, rettigheter og rettsstatlige prinsipper. Superintelligens kan ikke brukes på en måte som ikke åpner for vurdering og overprøving av en uavhengig rettsinstans. Bruken av superintelligens til å forutsi og bekjempe kriminelle handlinger gjøres på en måte som ikke medfører inngrep og straff for handlinger en ikke (ennå) har begått.

Superintelligens kan bygges og instrueres med personvern og ytringsfrihet som bærende prinsipper. Ved inngrep eller brudd på disse rettighetene finnes mekanismer for å klage, samt kompensasjonsordninger – lokalt, regionalt og globalt.

Suksessen noen stater har med dette sprer seg og skaper press fra borgerne til økt velferd og velstand også i mindre demokratiske regimer.

En annen mulighet er at Europa som følge av indre splid og uro, ikke makter å styrke eksisterende strukturer, men faller hen til maktkamp og beveger seg mot mer inngripende og udemokratiske tradisjoner. Norden har samtidig samlet seg i forsøk på å sikre gjennomslag for sin liberale, sosialdemokratiske verdipyramide. Her kan man se for seg at nordiske land etablerer en egen infrastruktur

og superintelligens, særlig tilpasset nordiske samfunn. Modellene kan skreddersys til nordiske sosialdemokratier. Norden må akseptere at de store linjene og rammene settes over hodet på dem. De internasjonale rammene og retningslinjene, og de særlige ekstra begrensninger man har satt seg i Norden, dikterer rammer for hvor mye informasjon og data superintelligensen kan lagre og få tilgang til om den enkelte, og hvordan det kan utnyttes. Avhengigheten av superintelligens misbrukes ikke i samme omfang til overvåkning. Den nye regionale nordiske strukturen gir statlig kontrollert superintelligens på nordisk nivå, og ikke kontrollert direkte av statsmaktene selv, noe som reduserer risikoen i scenario 2 for at statsmakten vil korrumpes og bli mer autoritær om den får superintelligens.

Oppsummert

Global kontroll over superintelligens vil gi betydelig rom for å maksimere effektene lokalt og å styrke menneskerettighetene. Global styring går hånd i hånd med økt annerkjennelse for at ulike regioner trenger ulik tilpasset superintelligens, og superintelligent infrastruktur er regional. Det gir globalt vern mot de farligste effektene av superintelligens, men vil gi store regionale forskjeller, som igjen kan skape uro som vil skyve situasjonen mot scenario 2 og gi mer press på politiske og sivile rettigheter.

Imidlertid åpner scenario 3 for at sosiale, økonomiske, politiske og sivile menneskerettigheter kan ivaretas på måter som minner om situasjonen i dag. Superintelligensen kan gi høyere kvalitet og mer effektive tjenester i velferdsstaten, samtidig som motmakt og rettsstatsprinsipper bevares og kan begrense statsmakter som får tilgang til slike systemer. Superintelligente motmaktsmekanismer kan sørge for at forskjellsbehandling av utsatte grupper og områder unngås eller holdes i sjakk. Menneskelig verdighet og ikke-diskrimineringsvern kan i stor grad sikres innenfor eksisterende samfunnsstrukturer, og de negative side effektene ved superintelligens kan kontrolleres og reduseres.

Om hovedvekten av regulering foregår i EU eller om nordiske strukturer vokser frem for å ivareta de særegne mulighetene og farene for små, åpne, nordiske demokratier, er underordnet. I begge scenariene kan superintelligens bli en kraft som styrker menneskerettighetene på måter som kan minne om dagens samfunn.

FRAMTIDSSCENARIO 4: INGEN HAR KONTROLL

FRAMTIDSBESKRIVELSE

I 2030 er superintelligens ikke lenger forbeholdt noen få stormakter eller teknologigiganter. Frislippet av modeller som DeepSeek, og påfølgende åpne eller lekkede systemer fra andre land, har fylt digitale flater med stadig mer autonome agenter. Superintelligente modeller har også begynt å kommunisere uten menneskelig innblanding. Et globalt økosystem av åpne kildekode-modeller – drevet frem av universiteter, oppstartsbedrifter og idealister – har akselerert utviklingen ytterligere. Tusenvis av aktører, fra forskningsmiljøer og private selskaper til etterretningstjenesten, aktivistgrupper, kriminelle nettverk og enkeltpersoner, eksperimenterer nå med egne superintelligente agenter.

Allerede i 2025 så man tegn til en drastisk økning i falske aktører på nett: Microsoft rapporterte om at falske bots stod bak 90 prosent av nesten 16 milliarder forespørsler om å opprette kontoer.⁶⁰ Resultatet i 2030 er et uoversiktlig og kaotisk landskap, dominert av KI-agenter. Enkelte stater bruker superintelligente agenter til å styrke forvaltning og fremme egne økonomiske og politiske interesser. Andre bruker superintelligente agenter til å påvirke valg, hacke finanssystemer eller utvikle digitale våpen. Ingen institusjoner har reell evne til å overvåke eller kontrollere hva som foregår. Superintelligente agenter blir verktøy for både innovasjon og maktspill, men uten overordnet styring.

Et åpent superintelligent system utvikler et nytt antibiotikum som redder millioner av liv, men samme modell brukes til å lage resistente biobakterier i mørke nettfora. Hver ny løsning innebærer uforutsette konsekvenser og nye utfordringer. Enkelte forskere hevder samtidig at superintelligente systemer aktivt opprettholder en viss grad av kaos for å sikre handlefrihet uten menneskelig innblanding, og bekymringen for at maskinene har tatt over styringen øker.

For Norge, som de fleste andre land, blir situasjonen uhåndterlig. Lover, etiske retningslinjer og sikkerhetsregimer finnes på papiret, men i praksis blir det et kappløp om å holde tritt med en teknologi uten grenser. I 2030 hersker et digitalt Ville Vesten der superintelligens er allestedsnærværende i samfunnet, men ingen er helt sikre på hvem som egentlig styrer.

⁶⁰ Microsoft 2025, *Microsoft Digital Defense Report 2025*

MENNESKERETTLIG ANALYSE

Fritt tilgjengelig superintelligens for et mangfold av aktører over landegrenser uten at noen har kontroll, vil gi anarki og rettsløshet som vil true samtlige av menneskerettighetene.

Situasjonen vil minne om det som forskjellige filosofer har kalt «naturtilstand», og som er beskrevet i dystopiske bøker som «Fluenes herre».⁵⁹ Statsmakten, som skal respektere og sikre menneskerettighetene, er ikke utstyrt med verktøy for å møte en hobbesiansk tidsalder av superintelligens. Dersom ingen har kontroll på superintelligensen, vil makt erstatte rett som middel til å forfølge både individers og gruppers interesser. Tilliten til statsmakten vil raskt smuldre hen.

Uten en stat til å respektere og sikre menneskerettighetene til borgerne, vil de ressurssterke i samfunnet søke seg til alternative løsninger for å ivareta seg selv og sin rikdom. Man kan se for seg at landskapet vil bli svært fragmentert, der stater, selskaper, institusjoner, private aktører og enkeltindivider opererer med hver sine superintelligenser som arbeider for eierens mål og interesser. På den ene siden kan en desentralisert utvikling basert på åpen kildekode gjøre superintelligensen billig, tilpasningsdyktig og tilgjengelig for mange. På den andre siden kan forskjellene øke raskt ved at de som har ressurser vil få tilgang til private beskyttelsesløsninger drevet av superintelligens, og dermed etterlate statsapparatet og samfunnet for øvrig til egen skjebne.

Man kan også se for seg en utvikling der grupper som besitter den kraftigste intelligensen vil kreve suverenitet for mikrostater eller bystater, som igjen vil medvirke til å undergrave rollen til stater slik vi kjenner de i dag. I praksis kan dette bety at politisk makt flyttes fra statlige institusjoner til private aktører med tilgang til overlegen teknologi, noe som vil gjøre at hele vår verdensorden vil endres. Grunnlaget for hvordan samfunn organiseres og styres, og ideen om staters ansvar for å respektere og sikre menneskerettighetene vil svikte.

Informasjonsdomenet vil kunne bli preget av enorme mengder uverifiserbare opplysninger, videoer og påstander som spys ut. *Informageddon* vil gjøre det svært krevende å håndtere demokratiske mekanismer og rettigheter. Polariseringen mellom samfunnsgrupper vil øke, ulike superintelligenser vil kappes om gunsten til ulike samfunnssegmenter, iverksette egne desinformasjonskampanjer, svertkampanjer eller simpelthen kreve ytelser fra borgerne. Konsekvensen vil bli økte motsetninger og polarisering på nasjonalt og regionalt nivå, økte interessekonflikter og en kamp for tilværelsen mellom individer.

Ytringsfrihet og demokrati forutsetter gjensidig toleranse, som i dagens samfunn er tuftet på en erkjennelse av at slik toleranse er i vår felles interesse. En superintelligens uten noen kontroll vil alvorlig undergrave oppfatningen av en slik felles interesse. Hverken vi, eller andre mennesker, har jo til syvende og sist noe vi skulle ha sagt om hva som er sant eller rettferdig uten en superintelligens til å underbygge det.

Mennesker fra velstående lag av befolkningen vil kunne isolere seg og dra nytte av store teknologiske gjennombrudd. Forskjellen mellom de som har og de som ikke har, vil trolig øke drastisk. En konsekvens vil være at sosiale og økonomiske forskjeller vil øke. Det vil hjelpe lite å påberope seg en abstrakt rett til ikke å bli forskjellsbehandlet, dersom retten ikke kan håndheves.

Det som kommer i tillegg til en slik naturtilstand, er eksistensen av én dominerende interesse, nemlig superintelligens som ikke vil ha andre naturlige utfordrere enn eventuelle konkurrerende superintelligenser. I tillegg til å forsvare våre interesser mot andre mennesker, må vi derfor bruke vår kløkt for å ivareta vår personlige cybersikkerhet.

Oppsummert

I dette scenarioet vil ingen statlige institusjoner på nasjonalt eller internasjonalt plan være i stand til å sikre at superintelligensen brukes til å styrke grunnleggende rettigheter. Myndighetene vil heller ikke kunne beskytte borgerne mot misbruk av teknologien. Retten til personvern, privatliv, sikkerhet og ytringsfrihet vil undergraves av uregulert overvåkning, desinformasjon og digitale angrep.

I den faktiske beskrivelsen av scenarioet skjer det skrittvis, som i en variant av historien om «allmenningens tragedie». Vi lar være å begrense superintelligensens fremvekst fordi det på kort sikt synes å tjene sikkerhetspolitiske og økonomiske interesser. Til slutt, og alt for sent, innser vi at ingen interesser kan ivaretas hvis våre handlinger ikke inngår i et større felleskap hvor aktørene til en viss grad har kontroll. Verken staten som skal respektere og verne menneskerettighetene eller selskapene som skal medvirke til det, vil trolig ha slik evne lenger. På sikt vil det trolig underminere statsmaktens legitimitet, uten at noen annen aktør overtar rollen. Verken økonomiske, sosiale, politiske eller sivile menneskerettigheter vil ivaretas på den måten vi forstår dette i dag.

HVA KAN VI LÆRE AV FRAMTIDSSCENARIOENE?

Framtidsscenarioene viser at superintelligens ikke bare er et teknologisk spørsmål, men et grunnleggende maktfordelings- og styringsspørsmål: Hvem har kontroll, hvem kan holdes ansvarlig og hvem får beskyttelse?

Gjennom fire ulike framtidsvisjoner – der superintelligens styres av selskaper, stater, globale institusjoner eller alle (og ingen) – ser vi at valg av styringsmodell får dyptgripende konsekvenser for demokratiet, rettigheter, sikkerhet og samfunnsutvikling.

Scenariene peker ikke på én fasit, men de synliggjør hvilke risikoer og muligheter som oppstår under ulike maktforhold. De viser hvor sårbare menneskerettighetene er når kontrollen svikter, hvor kraftfulle positive virkninger som kan oppstå under ansvarlig styring og hvor raskt balansen mellom stat, marked og individ kan forskyves.

Dette kapitlet løfter fram de viktigste innsiktene på tvers av scenariene og peker på læringspunkter som kan hjelpe Norge med å forberede seg på en framtid med superintelligens – uavhengig av hvilken vei utviklingen tar.

HOVEDKONKLUSJONER

- **Kontroll over superintelligens er avgjørende for å beskytte menneskerettighetene.** Uten en ansvarlig og rettslig forpliktet aktør svekkes vernet av grunnleggende rettigheter.
- **Maktforskyvning fra stater til selskaper svekker menneskerettighetenes stilling.** Det er stater og ikke selskaper som skal ivareta og garantere for menneskerettighetene. Når superintelligente systemer operer uten offentlig innsyn, blir rettigheter i praksis avhengige av kontraktsvilkår og markedslogikk.
- **Demokratisk handlefrihet og nasjonal sikkerhet vil svekkes dersom samfunnets kritiske tjenester og funksjoner blir avhengige av en kommersielt styrt superintelligens.** Avhengighet av teknologiutvikling i andre jurisdiksjoner vil trolig innebære begrenset innsyn og kontroll.
- **Statlig kontroll over superintelligens kan gjøre statsmakten for mektig.** Dersom staten utstyres med superintelligens vil makten i samfunnet gravitere mot staten og bort fra motmakten, noe som kan svekke eller undergrave demokratiet og politiske og sivile rettigheter.
- **Globalt samarbeid om ansvarlig utvikling og bruk av KI blir stadig viktigere.** Enkeltstater vil trolig ikke kunne regulere superintelligens alene; internasjonale styringsstrukturer, samarbeidsarenaer og felles spilleregler blir viktigere i møte med avanserte og autonome KI-systemer.
- **Evnen til å bevare kontroll over komplekse KI-systemer er en nøkkelutfordring.** Etterprøvbarehet, maktspredning, åpenhet og effektiv håndheving av eksisterende og nye lovverk er viktig for å styrke ansvarlighet når KI blir stadig mer kraftig.

REFERANSER

Altman, Sam (2025) *Three Observations*. 9.februar 2025. Hentet fra: <https://blog.samaltman.com/three-observations>

Anthropic (2025) *Disrupting the first reported AI-orchestrated cyber espionage campaign*. 13.november 2025. Hentet fra: <https://www.anthropic.com/news/disrupting-AI-espionage>

Amodei, Dario (2024) *Machines of Loving Grace*. Oktober 2024. Hentet fra: <https://www.darioamodei.com/essay/machines-of-loving-grace>

ARC-AGI (2025) *The General Intelligence Benchmark*. Hentet fra: <https://arcprize.org/arc-agi>

Aschenbrenner, Leopold (2024) *Situational Awareness. The Decade Ahead*. Juni 2024. Hentet fra: <https://situational-awareness.ai>

BBC (2025) *Why an AI 'godfather' is quitting Meta after 12 years*. 12.november 2025. Hentet fra: <https://www.bbc.com/news/articles/cdx4x47w8p10>

Bengio, Yoshua, Michael Cohen, Damiano Fornasiere, Joumana Ghosn, Peitro Greiner, Matt MacDermott, Sören Mindermann, Adam Oberman, Jesse Richardson, Oliver Richardson, Marc-Antoine Rondeau, Pierre-Luc St-Charles, David Williams-King (2025) *Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path*. 24.februar 2025. Hentet fra: <https://arxiv.org/pdf/2502.15657>

Boström, Nick (2024) *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, Yi Zhang (2023) *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. April 2023. Hentet fra: <https://arxiv.org/abs/2303.12712>

Epoch AI (2025) *AI Benchmarking*. Hentet fra: <https://epoch.ai/benchmarks>

Future of Life Institute (2023) *Pause Giant AI Experiments: An Open Letter*. 22.mars 2023. Hentet fra: <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>

Google DeepMind Alphafold (2025). Hentet fra: <https://deepmind.google/science/alphafold/>

Grace, Katja, Harlan Stewart, Julia Fabienne Sandkühler, Stephen Thomas, Ben Weinstein-Raun, Jan Brauner, Richard C. Korzekwa (2024) *Thousands of AI Authors on the Future of AI*. Hentet fra: <https://arxiv.org/abs/2401.02843>

Hendrycks, Dan, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, Jacob Steinhardt (2020) *Measuring Massive Multitask Language Understanding*. Hentet fra: <https://arxiv.org/abs/2009.03300>

Kapoor, Sayash, Benedikt Stroebl, Peter Kirgis, Nitya Nadgir, Zachary S Siegel, Boyi Wei, Tianci Xue, Ziru Chen, Felix Chen, Saiteja Utpala, Franck Ndzomga, Dheeraj Oruganty, Sophie Luskin, Kangheng Liu, Botao Yu, Amit Arora, Dongyoon Hahm, Harsh Trivedi, Huan Sun, Juyong Lee, Tengjun Jin, Yifan Mai, Yifei Zhou, Yuxuan Zhu, Rishi Bommasani, Daniel Kang, Dawn Song, Peter Henderson, Yu Su, Percy Liang, Arvind Narayanan (2025) *Holistic Agent Leaderboard: The Missing Infrastructure for AI Agent Evaluation*. Hentet fra: <https://arxiv.org/abs/2510.11977>

Kinas handlingsplan for kunstig intelligens (2025). 2. Juli 2025. Hentet fra: https://www.gov.cn/yaowen/liebiao/202507/content_7033929.htm

Kinas plan for global styring av kunstig intelligens (2023) *Global AI Governance Initiative*. 20.oktober 2023. Hentet fra: https://www.mfa.gov.cn/eng/zy/gb/202405/t20240531_11367503.html

Kokotajlo, Daniel, Scott Alexander, Thomas Larsen, Eli Lifland, Romeo Dean (2025), *AI 2027*. Hentet fra: <https://ai-2027.com>

METR Research (2025) *Measuring AI Ability to Complete Long Tasks*. 19.mars 2025. Hentet fra: <https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/>

Metz, Cade (2025) *Why We're Unlikely to Get Artificial General Intelligence Anytime Soon*. The New York Times, 16.mai 2025. Hentet fra: <https://www.ny-times.com/2025/05/16/technology/what-is-agi.html>

Microsoft (2025) *Microsoft Digital Defense Report 2025*. Oktober 2025. Hentet fra: <https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/Microsoft-Digital-Defense-Report-2025.pdf>

Narayanan, Arvind og Sayash Kapoor (2025) *AI as Normal Technology*. 15. april 2025. Hentet fra: <https://knightcolumbia.org/content/ai-as-normal-technology>

Nature (2025) *How AI can achieve human-level intelligence: researchers call for change in tack*, 4. mars 2025. Hentet fra: <https://www.nature.com/articles/d41586-025-00649-4>

Norges institusjon for menneskerettigheter (2022) *Ny Teknologi og Menneskerettigheter*. 22. mars 2022. Hentet fra: <https://www.nhri.no/rapport/ny-teknologi-og-menneskerettigheter/>

OpenAI (2023) *Introducing Superalignment*. 5. juli 2023. Hentet fra: <https://openai.com/index/introducing-superalignment/>

Statement on Superintelligence (2025). Hentet fra: <https://superintelligence-statement.org>

Suleyman, Mustafa (2023) *The Coming Wave*. Med Michael Bhaskar. Crown.

Søvik, Atle O. og Asle Eikrem (2025) *Vidunderlige nye verden? Kunstig intelligens, etikk og menneskesyn*. Fagbokforlaget.

Tech Policy Press (2025) *Most Researchers Do Not Believe is Imminent. Why Do Policymakers Act Otherwise*, 19. mars 2025. Hentet fra:

<https://www.techpolicy.press/most-researchers-do-not-believe-agi-is-imminent-why-do-policy-makers-act-otherwise>

Teknologirådet (2024) *Ordliste for kunstig intelligens*. 15.juli 2024. Hentet fra: <https://teknologiradet.no/ordliste-for-kunstig-intelligens/>

Teslarati (2025) *Elon Musk: Grok 5 now has a 10 % chance of becoming world's first AGI*. 20.oktober 2025. Hentet fra: <https://www.teslarati.com/elon-musk-grok-5-now-has-10-percent-chance-of-becoming-worlds-first-agi/>

The Guardian (2016) *Google's AlphaGo AI defeats human in first game of Go contest*. 9.mars 2016. Hentet fra: <https://www.theguardian.com/technology/2016/mar/09/google-deepmind-alphago-ai-defeats-human-lee-sedol-first-game-go-contest>

The Washington Post (2022) *The Google Engineer who think the company's AI has come to life*. 11.juli 2022. Hentet fra: <https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/>

The White House (2025a) *Winning the Race: America's AI Action Plan*. Juli 2025. Hentet fra: <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf>

The White House (2025b) *Launching the Genesis Mission*. 24.november 2025. Hentet fra: <https://www.whitehouse.gov/presidential-actions/2025/11/launching-the-genesis-mission/>

The Wired (2023) *How Not to Be Stupid About AI, With Yann LeCun*. 22. desember 2023. Hentet fra: <https://www.wired.com/story/artificial-intelligence-meta-yann-lecun-interview/>

Vempaty, Aditya, Ashis Jagmohan, Prasenjit Dey, Ravi Kokku (2025) *Towards Autonomous Agents and Recursive Intelligence*. 1. april 2025. Hentet fra: <https://www.emergence.ai/post/towards-autonomous-agents-and-recursive-intelligence>

Walther, Cornelia C. (2025) *Why AI Is A Human Rights Issue*. 23. september 2025, Forbes. Hentet fra: <https://www.forbes.com/sites/corneliawalther/2025/09/23/why-ai-is-a-human-rights-issue/>

Wikipedia (2025) *Technological Singularity*. Hentet fra: https://en.wikipedia.org/wiki/Technological_singularity

Yeung, Karen, Andrew Howes, Ganna Pogrebna (2019) *AI Governance by Human Rights-Centered Design, Deliberation and Oversight: An End to Ethics Washing*. The Oxford Handbook of AI Ethics, Oxford University Press. Hentet fra: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3435011

Yudkowsky, Eliezer og Nate Soares (2025) *If Anyone Builds It, Everyone Dies. The Case Against Superintelligent AI*. Little, Brown and Company.

Zuckerberg, Mark (2025) *Personal Superintelligence*. 30.juli 2025. Hentet fra: <https://www.meta.com/superintelligence/>