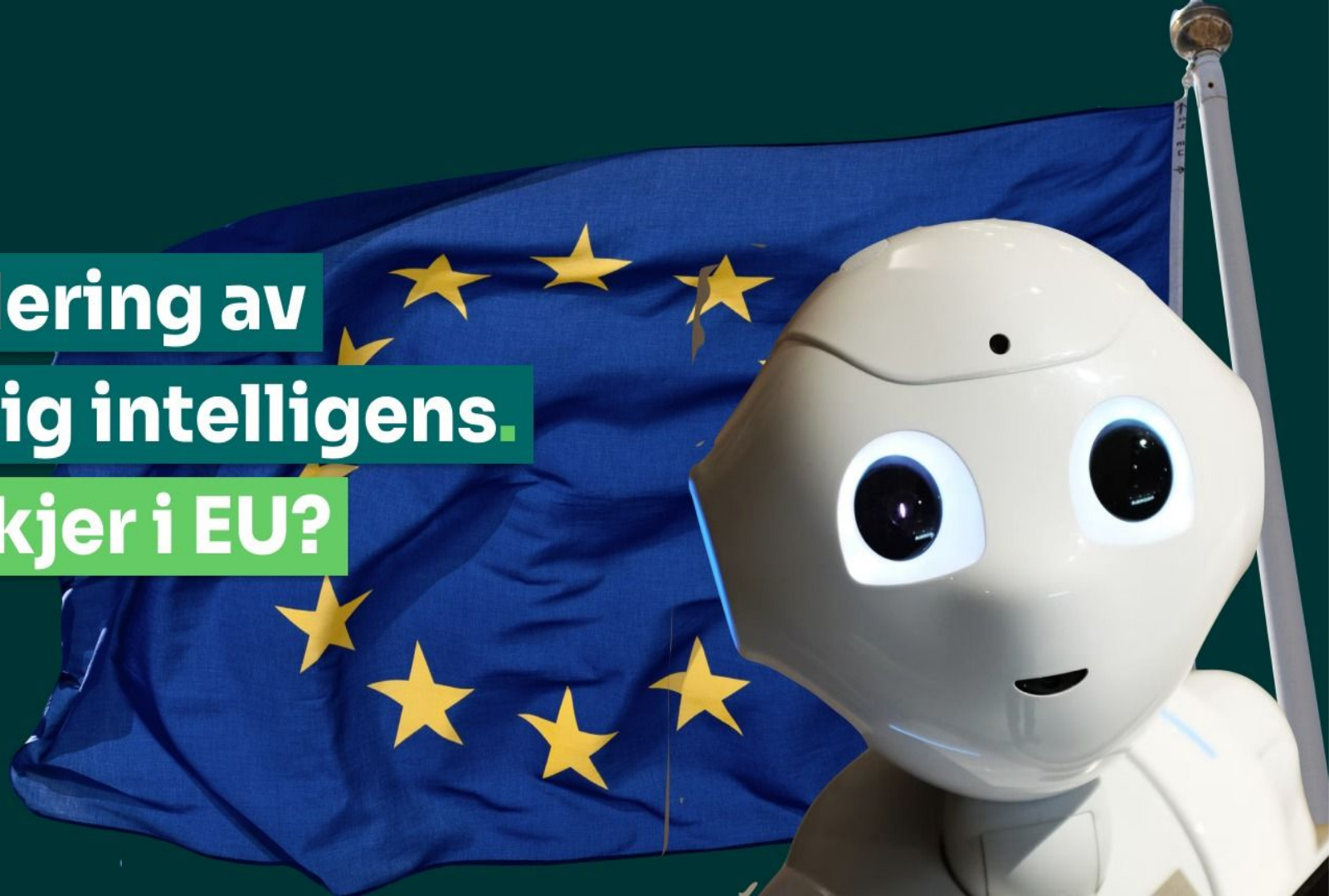
Regulering av kunstig intelligens. Hva skjer i EU?



VENSTRE

Hva er KI?

«The ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. Intelligent beings are those that can adapt to changing circumstances»

«The capability of a machine to imitate intelligent human behavior»

KI som definert i den foreslåtte reguleringen

«Software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with»

Hvordan brukes KI?

- Variert bruksområde, ikke lenger hypotetisk og fremtidig
- Chatbot, lånesøknad, netflix
- NAV utvikler nå en kunstig intelligens som kan predikere varigheten på individers sykefravær
- Varierende grad av alvorlighet og inngripen

Hvordan fungerer KI?

- KI basert på maskinlæring
  - Lærer basert på å finne mønstre datasett, framfor å programmeres til en gitt løsning
- Medfører utvikling på sikt → i motsetning til annen teknologi vil derfor kunstig intelligens oppføre seg annerledes over tid
- Begynner som et blankt ark og ender opp svært kompleks
  - Logikk, metode og beslutningsgrunnlag endrer seg derfor løpende
- En konsekvens av dette er at det er vanskelig å følge KIens logikk og fremgangsmåte

Positive virkninger av KI

- Økt effektivitet og presisjon
- Kan oppdage mønstre og sammenhenger mennesker ikke ville oppdaget
- Kan benyttes til effektivisering av for eksempel diagnostisering og saksbehandling

Utfordringer ved KI

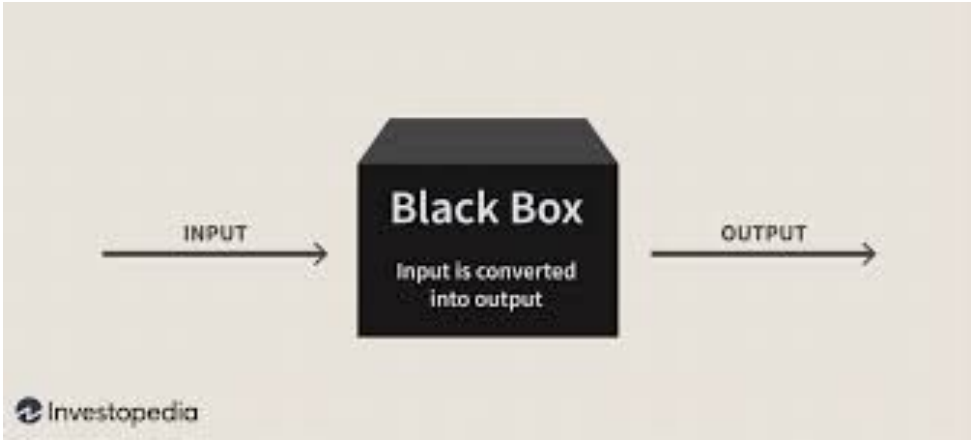
- Elon Musk og Stephen Hawking har advart mot negative konsekvenser av KI
- Utfordringer knyttet til forskjellen mellom et menneske og en KI
- KIen er begrenset til den informasjonen den har og oppdraget den er satt til å utføre
    - Tar ikke høyde for hvorvidt måten den løser oppdraget for eksempel er uetisk eller om informasjonen den baserer sin logikk på er diskriminerende eller rasistisk

Automation bias/automatiseringsskjevhet

- «Menneskelig tilbøyelighet til å favorisere informasjon fra automatiserte beslutningssystemer til fordel for annen informasjon, til tross for at den andre informasjonen er riktig»
- Eksempel radiologi - legen kan vurdere at hun tar feil, til tross for å sitte på fagkunnskap som overgår det automatiserte beslutningssystemet/Klens kompetanse
- Kan medføre at Klens forslag blir fulgt, til tross for å være et utslag feilvurderinger som burde ha bli oppdaget

Black box-problematikk

- KI-systemer som har en logikk som ikke er forklarbar av mennesker
- Logikken bak en beslutning kan derfor ikke forklares
- Vanskelig å oppdage om beslutningen for eksempel skyldes uetiske vurderinger



Eksempel - KI brukt i radiologi

- Beslutningsstøtte for radiologer
- KIen leser av røntgenbilder og sammenligner bildene med bilder fra både friske og syke → finner mønstre og oppdager avvik
- Forslagene fra KIen gjennomgås og det vurderes om disse skal tas til følge
- Mer presist og effektivt enn menneskelig/manuell gjennomgang av røntgenbildene
    - Går gjennom flere røntgenbilder på kortere tid
    - Oppdager lettere sykdomstegn og «flagger» derfor flere potensielle sykdomstilfeller
- Automation bias og black box-problematikk kan likevel skape utfordringer
    - Hva om legen stoler mer på Kien enn på sin egen vurdering?
    - Hva om manglende forståelse av KIens logikk gjør at man ikke oppdager feilvurderinger?

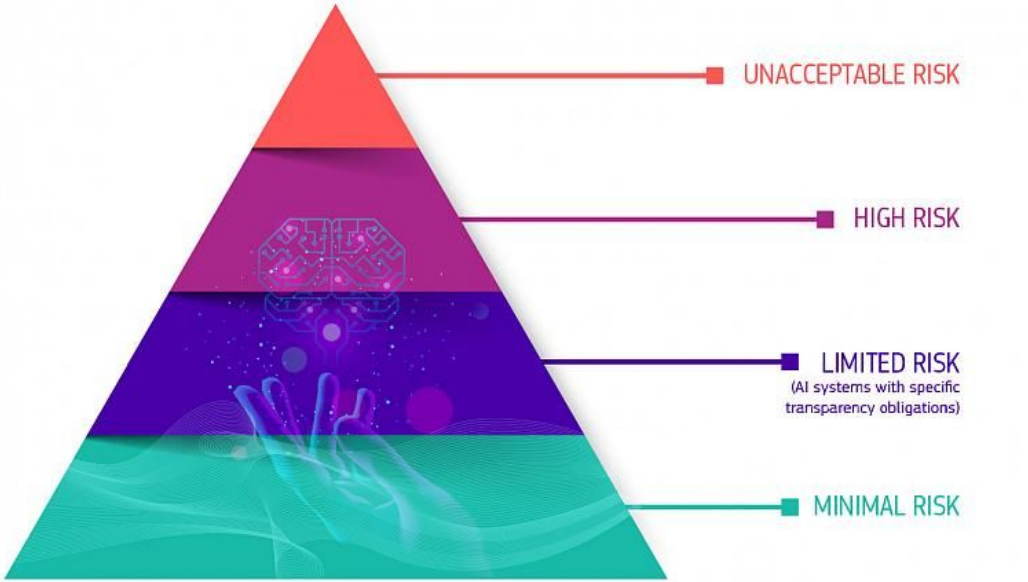
Introduksjon av foreslått KI-regulering

- Foreslått av EU-kommisjonen i 2021
- Formålet med reguleringen er å adressere og motvirke sikkerhetsutfordringer ved bruk av kunstig intelligens - som utfordringene nevnt over
- Enda ikke vedtatt, men anses som et uttrykk for EU-kommisjonens generelle tilnærming til KI
- Neste steg er vurdering i Europaparlamentet og Rådet for Den europeiske union
- Antagelig flere år til reguleringen vedtas og trer i kraft

Introduksjon av foreslått KI-regulering

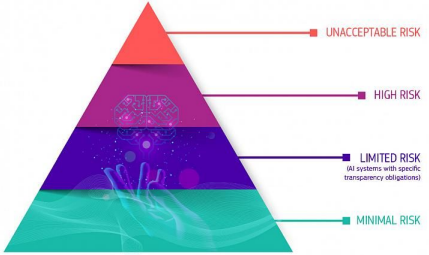
- Både rettet mot produsenter og brukere av KI
- Sektornøytral og skiller heller ikke mellom typen KI eller hvordan den fungerer
- Basert på risikonivåer som har stor konsekvens for det enkelte KI-systemet
- «Future proofing»
  - Generell og overordnet formulering
  - Ment å ramme både nåværende og fremtidig KI
  - Teksten i reguleringen er sjelden gjenstand for endring
  - Vedleggene oppdateres i takt med teknologisk utvikling
  - Hensyn til forutsigbarhet vs. endring
- Brudd på reglene sanksjoneres med bøter jf. art. 71
  - 30 000 000 EUR eller 6% av årlig omsetning
  - Det alternativet som gir høyest sum
  - Minner om sanksjoner etter GDPR

Risikobasert inndeling



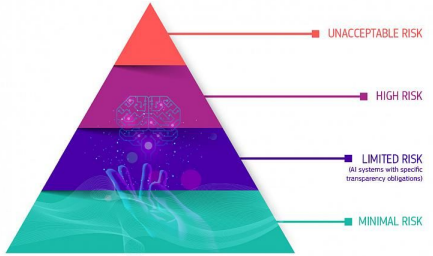
Unacceptable risk

- Høyeste nivået av risiko
- Disse KI-systemene er i utgangspunktet forbudt jf. artikkel 5
- Kan for eksempel være masseovervåking, sosiale scoring-systemer (som for eksempel foreslått i Kina),
- Vurdert å være i strid med grunnleggende EU-verdier for eksempel fordi de bryter med grunnleggende menneskerettigheter



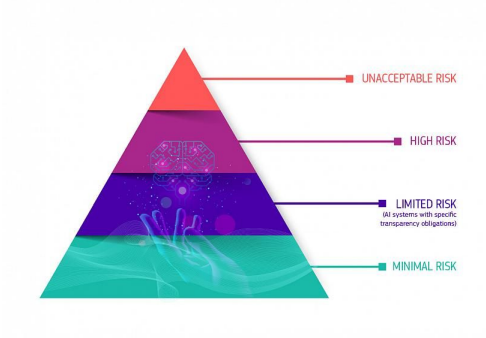
High risk

- Det høyeste tillatte risikonivået
- Vilkårene for hva som utgjør et high risk-system framgår av artikkel 6
- En rekke systemer faller inn under dette, blant annet KI som brukes til ansettelse og utdanning
- Omfatter systemer som har stor innvirkning på menneskers sikkerhet eller rettigheter
- Underlagt en rekke krav for å være tillatt
- Blant annet underlagt krav til menneskelig inngripen og conformity assessments



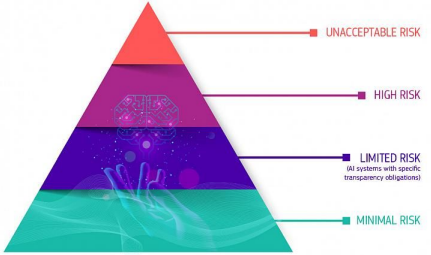
Limited risk

- Nest laveste risikonivå
- Omfatter KI-systemer som kan ha en viss risiko for manipulasjon
- For eksempel chatbots og deep fakes
- Stiller krav transparens
    - For eksempel opplysninger om at det er en KI man snakker med og ikke et menneske



Minimal risk

- Øvrige KI-systemer
  - Faller ikke inn i øvrige kategorier
- Stilles ingen krav til disse systemene i det foreslåtte regelverket
- Viktig å bemerke at disse systemene utgjør majoriteten av KI-systemene som er i bruk
- Vurdert slik at eksisterende lovgivning i tilstrekkelig grad adresserer og løser utfordringene ved disse systemene



Konsekvenser av risikoinndelingen

- Store forskjeller i krav som stilles til ulike risikonivå
  - Forutsetter klar grensedragning mellom risikonivåene
- Samspillet mellom selve reguleringen og vedleggene
  - Forutsetter kontinuerlig oppdatering

Artikkel 14 nr. 1:

«High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which the AI system is in use»

Konsekvenser av reguleringen

- Generell tilnærming gir reguleringen lenger levetid, men kan ha konsekvenser for forståelsen av krav som stilles
  - Hvilken innvirkning kan uklarhet ha for næringslivet?
  - Er det kun store selskaper som tør satse på KI, når det er en sjanse for at manglende etterlevelse av kravene kan bli dyrt?
- Sett i sammenheng med GDPR - er uklarhet et generelt problem?

Konsekvenser av reguleringen

- Burde det i større grad vært differensiert på type teknologi, eller burde definisjonen på KI som benyttes vært snevrere?
- Reguleringen er en konsekvens av avveining mellom ulike hensyn
  - For eksempel vern av grunnleggende rettigheter, personvernhensyn og insentiv til teknologisk utvikling
- Reguleringen kan ses som et uttrykk for EU-kommisjonens vekting mellom disse ulike hensynene
- Må antas at den videre prosessen vil medføre endringer