Kursus i algoritmiske grundprincipper og etik i AI: fra teori til praksis

Værktøjer til synkrone sessioner

CU1 | AI-etik - en praktisk tilgang

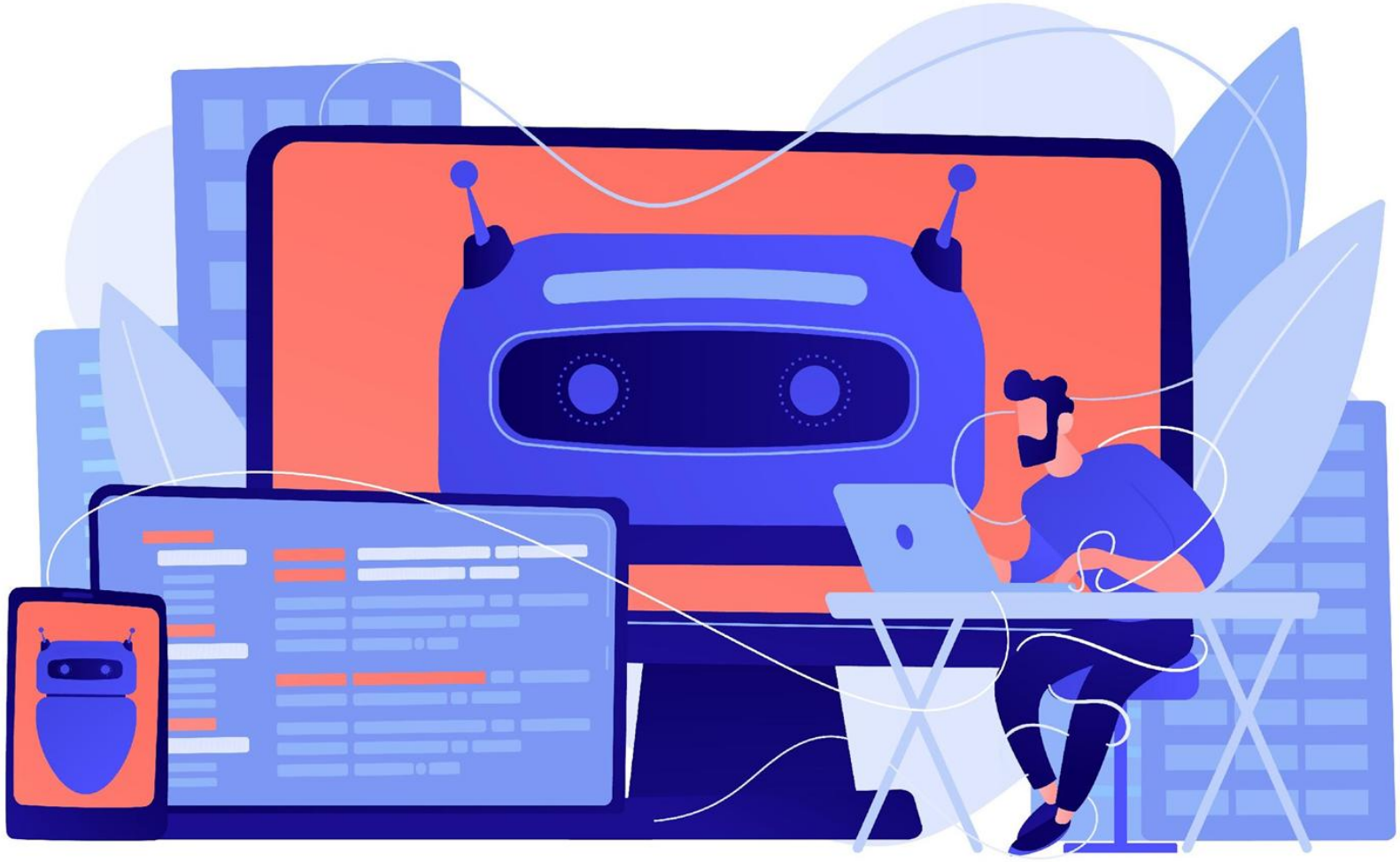
Understøttende PowerPoint-slides

Projektnummer | 2022-1-ES01-KA220-HED-000085257

INDHOLD

- INDLEDNING - 3
- ETISKE PRINCIPPER FOR AI - 8
- ETISKE RAMMER, RETNINGSLINJER OG VÆRKTØJER VEDR. AI - 23
- EU'S AI-FORORDNING - 33
- LIVSCYKLUS FOR AI-UDVIKLING - 40
- INDDRAGELSE AF INTERESSENTER I UDVIKLINGEN AF AI-LØSNINGER - 45
- ETISKE PRINCIPPER OG INDDRAGELSE AF INTERESSENTER I AI-UDVIKLINGENS LIVSCYKLUS - 53

INDLEDNING



BILLEDKILDE | Freepik

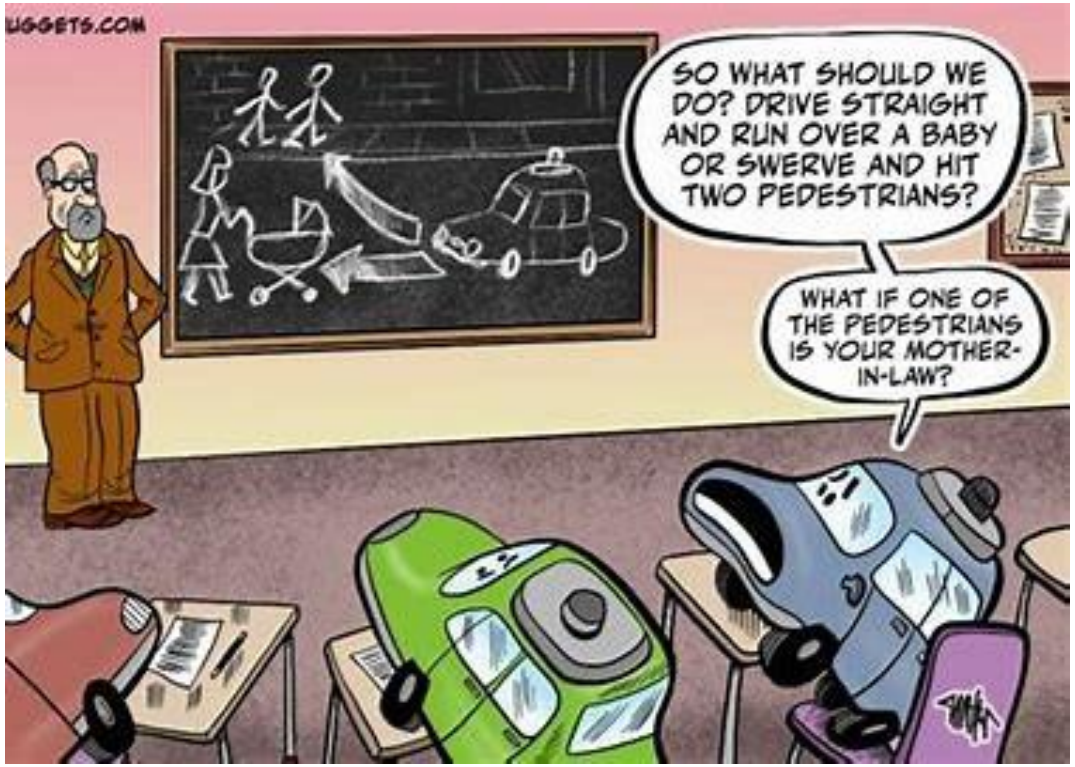
I DENNE KOMPETENCEENHED FINDER DU FØLGENDE EMNER:

- AI-etik og dens betydning for udvikling af AI-løsninger
- Etiske principper for AI
- Etiske rammer, retningslinjer og værktøjer
- Ekspertgruppe på højt niveau; troværdige AI-rammer
- EU's AI-forordning
- Livscyklus for AI-udvikling
- Inddragelse af interessenter i AI-udviklingens livscyklus
- AI-udviklingens livscyklus med tilhørende etiske principper og interessenter

VED AFSLUTNINGEN AF KOMPETENCEENHEDEN BØR DU VÆRE I STAND TIL AT:

- Kende til de etiske principper i udviklingen og implementeringen af AI
- Forstå forskellige etiske rammer og retningslinjer og deres anvendelse i virkelige scenarier, hvor AI-systemer indgår.
- Forstå vigtigheden af etiske principper og inddragelse af interessenter i udviklingsprocessen for etisk AI

Etiske overvejelser – hvad sker der, hvis...



BILLEDKILDE | Tegneserie: Undervisning i etik for AI - KDnuggets

- AI har ret til at afgøre, om du får et lån?
- AI skal afgøre, hvem der skal i fængsel?
- En selvkørende bil skal beslutte, om den vil redde en fodgængers liv eller livet for dem, der sidder i bilen?
- Et AI-baseret rekrutteringssystem kunne bestemme, om en kvindelig eller mandlig kandidat skulle ansættes som flykaptajn?

Kunstig intelligens revolutionerer livet

BILLEDKILDE | OECD; AI: Risk and Opportunity

Video fra: https://youtu.be/-CXkHs3cxa4

- AI påvirker alle aspekter af livet; det påvirker arbejde, fritid og fremsætter løsninger på globale problemer som klimaforandringer og adgang til sundhedspleje, samtidig med at det skaber betydelige udfordringer for regeringer og enkeltpersoner.
- Integrationen af AI i økonomi og samfund rejser spørgsmålet om, hvorvidt der er de politiske og institutionelle rammer, der er nødvendige for at styre AI's udvikling og anvendelse, så det bliver til gavn for samfundet.
- Omfattende etiske overvejelser har ført til oprettelsen og offentliggørelsen af forskellige tilgange til at sikre en etisk anvendelse af AI, som netop er emnet i dette kursus.

ETISKE PRINCIPPER FOR AI

BILLEDKILDE | Freepik

Etiske principper for AI

Hvad er de etiske principper for AI?

- **Etiske principper defineres som generelle retningslinjer, der tager højde for etiske problemer i forbindelse med AI-systemer.** Disse principper er ofte abstrakte og giver ikke specifikke instruktioner om, hvordan de skal anvendes i en bestemt sammenhæng.
- Etiske principper beskriver, hvad der genrelet forventes at blive betragtet som rigtigt og forkert samt andre etiske standarder. Etiske principper for AI henviser til normative begrænsninger for **"do's" og "don'ts"** i forbindelse med brug af algoritmer i samfundet.

(Prem 2023), (Zhou et al., 2020)

Etiske principper for AI

- Der er udgivet en lang række anbefalinger med hensyn til etisk anvendelse af kunstig intelligens.
- Anbefalingerne kommer fra forskellige sektorer som f.eks.:
  - OECD's anbefaling, som skitserer principper for AI-etik (offentliggjort maj 2019)
  - Beijing AI Principles for R&D (offentliggjort maj 2019)
  - Industrien/branchen (f.eks. Google, IBM, Microsoft, Intel)
  - Regeringer (f.eks. Montreal-erklæringen, Europa-Kommissionens ekspertgruppe)
  - Den akademiske verden (f.eks. Future of Life Institute, IEEE, AI4People)

(Prem 2023), (Morley et al 2020)

Etiske principper for AI

Eksempler på etiske principper

Illustrationen viser, hvordan forskellige rammer kombinerer de etiske principper, og hvordan de overlapper hinanden.
De vandrette pile forbinder emnerne med hinanden.

AI4People (udgivet november 2018) (Floridi et al. 2018)	Fem principper, der er afgørende for enhver etisk ramme for AI (L Floridi og Clement-Jones 2019)	Etiske retningslinjer for troværdig AI (offentliggjort april 2019) (Europa-Kommissionen 2019)	Anbefaling fra 'Council of Artificial Intelligence' (offentliggjort maj 2019) (OECD 2019b)
Gavnlighed	AI skal være til gavn for menneskeheden	Respekt for menneskelig autonomi	Inkluderende vækst, bæredygtig udvikling og velfærd
Ikke-skadelighed (*non-maleficence*)	AI må ikke krænke privatlivets fred eller underminere sikkerheden	Forebyggelse af skade	Robusthed, sikkerhed og tryghed
Autonomi	AI skal beskytte og styrke vores autonomi og evne til at træffe beslutninger og vælge mellem alternativer.		Menneskecentrerede værdier
Retfærdighed	AI skal fremme velstand	Retfærdighed	Retfærdighed
Forklarlighed	AI-systemer skal være forståelige og kunne forklares	Forklarlighed	Gennemsigtighed og forklarlighed Ansvarlighed

Modificeret fra Morley et al 2020

Etiske principper for AI

Fælles principper i dokumenter om etisk AI

BILLEDKILDE | Billedet er udviklet af VAMK

Etiske AI-principper viser et betydeligt overlap, hvilket tyder på en konvergens mod et samlet sæt etiske retningslinjer.

- Gavnlig og respektfuld over for mennesker og miljø (**gavnlighed**).
- Robust og sikker (**ikke-skadelig**).
- Respekt for menneskelige værdier (**autonomi**).
- Fair (**retfærdighed**).
- Forklarlig, ansvarlig og forståelig **(forklarlighed**).

Morley et al (2020), Jobin et al (2019)

Etiske principper for AI

BILLEDKILDE | Freepik

Gavnlighed

Gavnlighed i AI omfatter fremme af menneskehedens ve og vel

Princippet om gavnlighed understreger, at AI skal gavne menneskeheden og vores jordklode, og fremhæver vigtigheden af at fremme trivsel og miljømæssig bæredygtighed.

(Floridi et al 2018)

Etiske principper for AI

BILLEDKILDE | Freepik

Eksempel på princippet om gavnlighed

Anbefalinger til sundhedsvæsenet

- Forestil dig et AI-drevet sundhedssystem, der anbefaler behandlingsplaner baseret på patientdata. Hvis AI'en prioriterer behandlingsmuligheder, der er mere rentable for sundhedsudbyderen, frem for dem, der er mest gavnlige for patienterne, vil den ikke kunne overholde princippet om gavnlighed. I dette scenarie prioriterer AI'ens anbefalinger økonomiske gevinster frem for at maksimere patienternes velbefindende.
- IBM's Watson for Oncology stod over for etiske udfordringer, da den anbefalede behandlinger, der ikke altid var i overensstemmelse med medicinske retningslinjer eller patienternes interesser, hvilket gav anledning til bekymring om prioritering af profit frem for gode patientresultater. For mere information om sagen, se dette link.

(Floridi et al. 2018)

Etiske principper for AI

BILLEDKILDE | Freepik

Ikke-skadelighed (*non-maleficence*)

Ikke-skadelighed i AI fokuserer på at undgå skade.

Ikke-skadelighed fokuserer på at forhindre skade, hvad enten det er fra tilsigtede menneskelige handlinger eller utilsigtet maskinel adfærd, herunder utilsigtet indflydelse på menneskelig adfærd. Det handler om at undgå negative resultater – uanset kilden.

(Floridi et al. 2018)

Etiske principper for AI

BILLEDKILDE | Freepik

Eksempel på princippet ikke-skadelighed

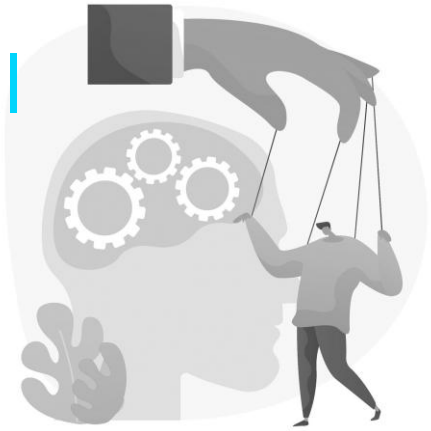
Forudsigende politiarbejde (*predictive policing*)

- Forestil dig en AI-algoritme, der bruges til forudsigende politiarbejde, og som i uforholdsmæssig grad er rettet mod minoritetsgrupper på grund af bias i de data, algoritmen blev trænet på. Dette AI-system kan utilsigtet videreføre skade og uretfærdighed ved at udsætte uskyldige personer for øget overvågning eller kontrol baseret på fejlagtige antagelser. På trods af intentionen om at reducere kriminalitetsraten kan AI'ens handlinger føre til uretfærdig chikane eller uretfærdig fokus på visse demografiske grupper.

- New Orleans' politis brug af *predictive policing*-softwaren "Palantir" var i uforholdsmæssig grad rettet mod minoritetssamfund, hvilket gav anledning til bekymring for, at systemisk bias og uretfærdighed i retshåndhævelsespraksis blev opretholdt. For mere information om sagen, se dette link.

Bias in algorithms – Artificial intelligence and discrimination (europa.eu)

Etiske principper for AI



BILLEDKILDE | Freepik

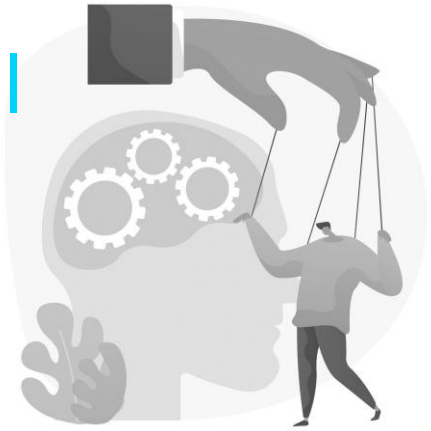
Autonomi

Autonomi i AI indebærer en omhyggelig balance mellem menneskelig og maskinel beslutningstagning og understreger behovet for at bevare menneskets selvbestemmelse.

- Autonomi indebærer en afbalancering af beslutningskompetencen mellem mennesker og AI. Den fastholder den iboende værdi af menneskelige valg i forbindelse med vigtige beslutninger og går ind for en "beslutte at uddelegere"-model, hvor mennesker bevarer den ultimative myndighed til at uddelegere og om nødvendigt tilsidesætte AI-beslutninger.

(Floridi et al. 2018)

Etiske principper for AI



BILLEDKILDE | Freepik

Eksempel på princippet om autonomi

Manipulation af sociale medier

- Et eksempel på AI, der underminerer autonomi, kunne være en anbefalingsalgoritme til sociale medier, der bruger personaliseret indhold til at manipulere brugernes adfærd og meninger uden deres udtrykkelige samtykke. Ved løbende at fodre brugerne med indhold, der stemmer overens med deres eksisterende overbevisninger eller fordomme, begrænser AI brugernes eksponering for forskellige perspektiver og påvirkninger og begrænser dermed deres autonomi til at træffe informerede beslutninger baseret på en bred vifte af oplysninger.
- **Et eksempel fra det virkelige liv:** Facebooks News Feed-algoritme, der er designet til at personalisere indhold baseret på brugernes interaktioner, er blevet kritiseret for at skabe filterbobler og ekkokamre, der potentielt forstærker sensationelt eller polariserende indhold og påvirker brugernes adfærd uden udtrykkeligt samtykke. For mere information, se dette link.

Bias in algorithms – Artificial intelligence and discrimination (europa.eu) / Electronic Frontier Foundation (EFF), Pew Research Center, or Data & Society

Etiske principper for AI

BILLEDKILDE | Freepik

Retfærdighed

Princippet om retfærdighed i AI handler om retfærdighed i beslutningstagningen og den samfundsmæssige fordeling, der sigter mod at mindske sociale forskelle og fremme retfærdighed.

- Princippet om retfærdighed handler om at bruge AI til at rette op på historiske uretfærdigheder, sikre en retfærdig fordeling af AI's fordele og forhindre nye skadelige virkninger eller disruption af den sociale solidaritet.

(Floridi et al. 2018)

Etiske principper for AI

BILLEDKILDE | Freepik

Eksempel på princippet om retfærdighed

Algoritmer til strafudmåling

- AI-drevne algoritmer, der bruges til strafudmåling, kan utilsigtet fastholde fordomme mod visse demografiske grupper. Hvis historiske data, der bruges til at træne disse algoritmer, afspejler forudindtagede beslutninger truffet af mennesker, kan AI-systemet anbefale længere straffe eller hårdere straffe til personer fra marginaliserede samfund, hvilket forværrer eksisterende uligheder i det strafferetlige system.
- COMPAS-systemet, der bruges til risikovurdering i forbindelse med strafudmåling, viste sig at udvise racemæssig bias, idet sorte tiltalte i uforholdsmæssig grad blev stemplet som havende en højere risiko for at begå kriminalitet, og systemet bidrog til uretfærdige strafudmålingsresultater. For mere information om sagen, besøg dette link.

Megan T. Stevenson og Christopher Slobogin, University of Pennsylvania Law Review

Etiske principper for AI

BILLEDKILDE | Freepik

Forklarlighed

Forklarlighed i AI betyder, at vi skal sikre os, at vi forstår, hvordan AI træffer sine beslutninger, og at vi er i stand til at forklare disse beslutninger tydeligt. Det sikrer, at AI-systemer er gennemsigtige og ansvarlige for deres handlinger.

(Floridi et al 2018)

Etiske principper for AI

BILLEDKILDE | Freepik

Eksempel på princip om forklarlighed

Kreditvurderingsalgoritmer

- Forestil dig et AI-drevet kreditscoringssystem, der bestemmer personers kreditværdighed, men mangler gennemsigtighed i beslutningsprocessen. Hvis ansøgere får afslag på lån eller tilbydes ugunstige vilkår uden at forstå de faktorer, som AI-algoritmen har taget i betragtning, underminerer det deres mulighed for at udfordre eller anke beslutningerne. Uden klare forklaringer på, hvordan AI'en vurderer kreditrisiko, kan enkeltpersoner føle sig uretfærdigt behandlet og miste tilliden til systemets integritet.

- Databruddet hos Equifax i 2017 afslørede millioner af forbrugeres følsomme oplysninger og gav anledning til bekymring over den manglende gennemsigtighed og sikkerhed i kreditvurderingsalgoritmerne, hvilket satte fokus på generelle brancheproblemer vedrørende forbrugernes begrænsede indsigt i faktorer, der påvirker kreditværdigheden. For mere information om sagen, besøg dette link.

Keith Frankish & William M. Ramsey

ETISKE RAMMER, RETNINGSLINJER OG VÆRKTØJER VEDR. AI

BILLEDKILDE | Freepik

Etiske rammer, retningslinjer og værktøjer vedr. AI

Etiske rammer

- **Etiske rammer er mere end principielle erklæringer på højt niveau. De giver en struktureret tilgang til etisk beslutningstagning.**
- Mange rammer for etisk AI har til formål at identificere potentielle etiske udfordringer og give nogle løsninger til at overvinde disse udfordringer eller mindske de tilhørende risici.
- En etisk ramme for AI fungerer som en grundlæggende struktur for udformning af love, regler, tekniske standarder og bedste praksis på tværs af forskellige sektorer og regioner.
- Typiske komponenter i etiske rammer er:
    - Begreber - relevante for at debattere de etiske aspekter
    - Princip - etiske principper (f.eks. værdier)
    - Bekymringer/betænkeligheder - hvordan principper trues af brugen og udviklingen af AI-systemer
    - Løsning - strategier, regler og retningslinjer for at løse problemet

(Floridi & Cowls, 2019), (Prem, 2023: 701) (Ayling & Chapman, 2022)

Etiske rammer, retningslinjer og værktøjer vedr. AI

Etiske værktøjer og retningslinjer

- **Etiske retningslinjer giver en praktisk vejledning til og specifikke regler for etisk adfærd i forskellige sammenhænge.**
- Etiske **værktøjer og retningslinjer** former AI-baseret innovation og støtter den **praktiske anvendelse af etiske principper for AI.**
- Værktøjer og retningslinjer forklarer, hvordan man anvender etiske principper i design, implementering og udrulning af AI.
- Rammer, retningslinjer og værktøjer overlapper ofte hinanden i deres brug.
- På de følgende sider præsenteres to eksempler på etiske rammer mere detaljeret: HLEG Trustworthy AI Framework og OECD AI Classification Framework.

(Zhou et al., 2020)

Etiske rammer, retningslinjer og værktøjer vedr. AI

Ekspertgruppe på højt niveau om kunstig intelligens; troværdig AI

- Selv om der er enighed om, at AI bør være etisk, er der debat om, hvad der udgør "etisk AI", og hvilke etiske krav og tekniske standarder der er nødvendige for at opnå det.
- Europa-Kommissionen præsenterede en strategi for udvikling af kunstig intelligens i Europa, som lægger vægt på etisk, sikker og avanceret kunstig intelligens.
- De etiske retningslinjer for pålidelig AI blev præsenteret af AI-HLEG i april 2019 for at fremme pålidelig AI.
- **Troværdig** AI er anerkendt som det vigtigste mål for at fremme tilliden til udvikling og anvendelse af AI, baseret på en robust ramme, der sikrer dens troværdighed.
- Se mere fra https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

(Leijnen et al., 2020), HLEG (2019). High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. Brussels: European Commission

Etiske rammer, retningslinjer og værktøjer vedr. AI

Troværdige AI-rammer

Troværdig AI har tre komponenter, som skal opfyldes gennem hele systemets livscyklus:

- Lovlig - respekterer alle gældende love og regler
- Etisk - respekterer etiske principper og værdier
- Robust - både fra et teknisk perspektiv og under hensyntagen til det sociale miljø

Ethics guidelines for trustworthy AI. (2019). Publications Office of the European Union.

Etiske rammer, retningslinjer og værktøjer vedr. AI

Ekspertgruppe på højt niveau om kunstig intelligens: rammer for troværdig AI

Grundlaget for troværdig AI

- Fire **etiske principper** baseret på grundlæggende rettigheder

Realisering af troværdig AI

- Implementerer syv **vigtige krav**

Vurdering af troværdig AI

- Til operationalisering af de vigtigste krav skal der skræddersys en **vurderingsliste** til den specifikke AI-applikation

HLEG (2019). High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. Brussels: European Commission.

Etiske rammer, retningslinjer og værktøjer vedr. AI

Grundlaget for troværdig AI: Etiske principper

Respekt for menneskelig autonomi

AI skal fremme selvbestemmelse, undgå manipulation og understøtte menneskets kognitive, sociale og kulturelle færdigheder. Det skal give mulighed for menneskelig tilsyn og meningsfulde valg, og støtte mennesker på arbejdspladsen.

Forebyggelse af skadelige virkninger

AI må ikke forårsage skade, AI skal beskytte den menneskelige værdighed, være sikker og robust og tage særligt hensyn til sårbare grupper og miljøet. Den skal forhindre negative virkninger på grund af magt- eller informationsubalance.

Retfærdighed

AI skal sikre en retfærdig fordeling af fordele og ulemper, forhindre forudindtagethed og diskrimination og fremme lige muligheder. Det kræver balance mellem midler og mål, der skal gives mulighed for at anfægte og klage over AI-beslutninger.

Forklarlighed

- AI-processer skal være gennemsigtige med klar kommunikation af muligheder og formål. Når fuld forklarlighed ikke er mulig, skal andre foranstaltninger som sporbarhed og reviderbarhed sikre, at rettighederne respekteres.

- AI HLEG opstiller fire etiske principper, der er forankret i grundlæggende rettigheder, og som skal respekteres for at sikre, at AI-systemer udvikles, implementeres og bruges på en troværdig måde.
- Etiske principper er specificeret som etiske imperativer, således at AI-udøvere altid bør stræbe efter at overholde dem.
- AI HLEG anerkender og håndterer også de potentielle spændinger mellem disse principper.

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Etiske rammer, retningslinjer og værktøjer vedr. AI

Realisering af troværdig AI: Vigtige krav

- De fastlagte etiske principper skal omdannes til endelige **krav** for realiseringen af troværdig AI. Disse krav gælder for alle deltagere i AI-systemets livscyklus:
- Udviklere: personer eller teams, der arbejder med AI-forskning, -design eller -udvikling.
- Udbydere: organisationer, der implementerer AI og tilbyder AI-drevne produkter eller services.
- Slutbrugere: Mennesker, der interagerer med AI-systemer, enten direkte eller indirekte.
- Det bredere samfund: dem, der påvirkes af AI, både direkte og indirekte.

Menneskelig udførelse og kontrol

Grundlæggende rettigheder, menneskelig udførelse og menneskelig kontrol

Teknisk robusthed og sikkerhed

Modstandsdygtighed over for angreb, sikkerhed, fall back-plan og generel sikkerhed, nøjagtighed, pålidelighed og reproducerbarhed

Beskyttelse af personlige oplysninger og datastyring

Respekt for privatlivets fred, datakvalitet og -integritet samt adgang til data

Gennemsigtighed

Sporbarhed, forklarlighed og kommunikation

Diversitet, ikke-diskrimination og retfærdighed

Undgåelse af urimelig bias, tilgængelighed, universelt design samt inddragelse af interessenter

Miljø- og samfundsmæssig velfærd

Bæredygtighed og miljøvenlighed, social indvirkning, samfund og demokrati

Ansvarlighed

Reviderbarhed, minimering og rapportering af negative virkninger, afvejninger og klageadgang.

HLEG (2019). High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. Brussels: European Commission

Etiske rammer, retningslinjer og værktøjer vedr. AI

Vurdering af pålidelig AI: ALTAI-vurderingslisten

- For at hjælpe med at udvikle troværdig AI er der udviklet et værktøj, som omsætter disse AI-principper til praktiske foranstaltninger.
- Værktøjet er kendt som *Assessment List for Trustworthy* AI (ALTAI) og indeholder en omfattende og dynamisk tjekliste, der fungerer som en køreplan for udviklere og og dem, der implementerer AI, så de kan integrere disse principper i deres AI-applikationer.
- ALTAI letter denne proces ved at tilbyde et sæt konkrete trin til selvevaluering, så man derved kan sikre, at AI-teknologier udvikles på en måde, der maksimerer brugernes fordele og samtidig minimerer eksponeringen for unødvendige risici.
- Adgang til ALTAI-værktøj

BILLEDKILDE | ALTAI

(High-Level Expert Group on AI (AI HLEG), 2020)

Etiske rammer, retningslinjer og værktøjer vedr. AI

OECD's ramme for AI-klassificering

- OECD's AI-klassifikationsramme, der er skabt the *OECD.AI Network of Experts*, er designet til at hjælpe forskellige interessenter med at evaluere potentialet og risikoen ved forskellige AI-systemer.
- Dette værktøj letter udviklingen af AI-strategier og har til formål at bevare sammenhængen i forskellige politikker på internationalt plan.
- Rammen er baseret på OECD's AI-principper, der støtter værdier som retfærdighed, gennemsigtighed, sikkerhed og ansvarlighed, og som slår til lyd for en af den menneskelige indflydelse samt det internationale samarbejde.

BILLEDKILDE | OECD-rammeværk

https://www.oecd.org/digital/artificial-intelligence/

EU'S AI-FORORDNING

BILLEDKILDE | Freepik

EU’s forordning om kunstig intelligens

- AI-forordningen er et udkast til en EU-lov om kunstig intelligens - den første af sin slags i verden. Det forventes, at loven vil træde i kraft i 2026.
- Den gælder for udvikling, implementering og brug af AI i EU, eller når dette vil påvirke mennesker i EU.
- Europa-Kommissionen offentliggjorde forslaget til forordningen om kunstig intelligens den 21. april 2021. Loven er den første juridiske ramme nogensinde, der fastsætter harmoniserede regler for udvikling, markedsføring og brug af kunstig intelligens i EU.

BILLEDKILDE | The AI Act Explorer | EU's lov om kunstig intelligens

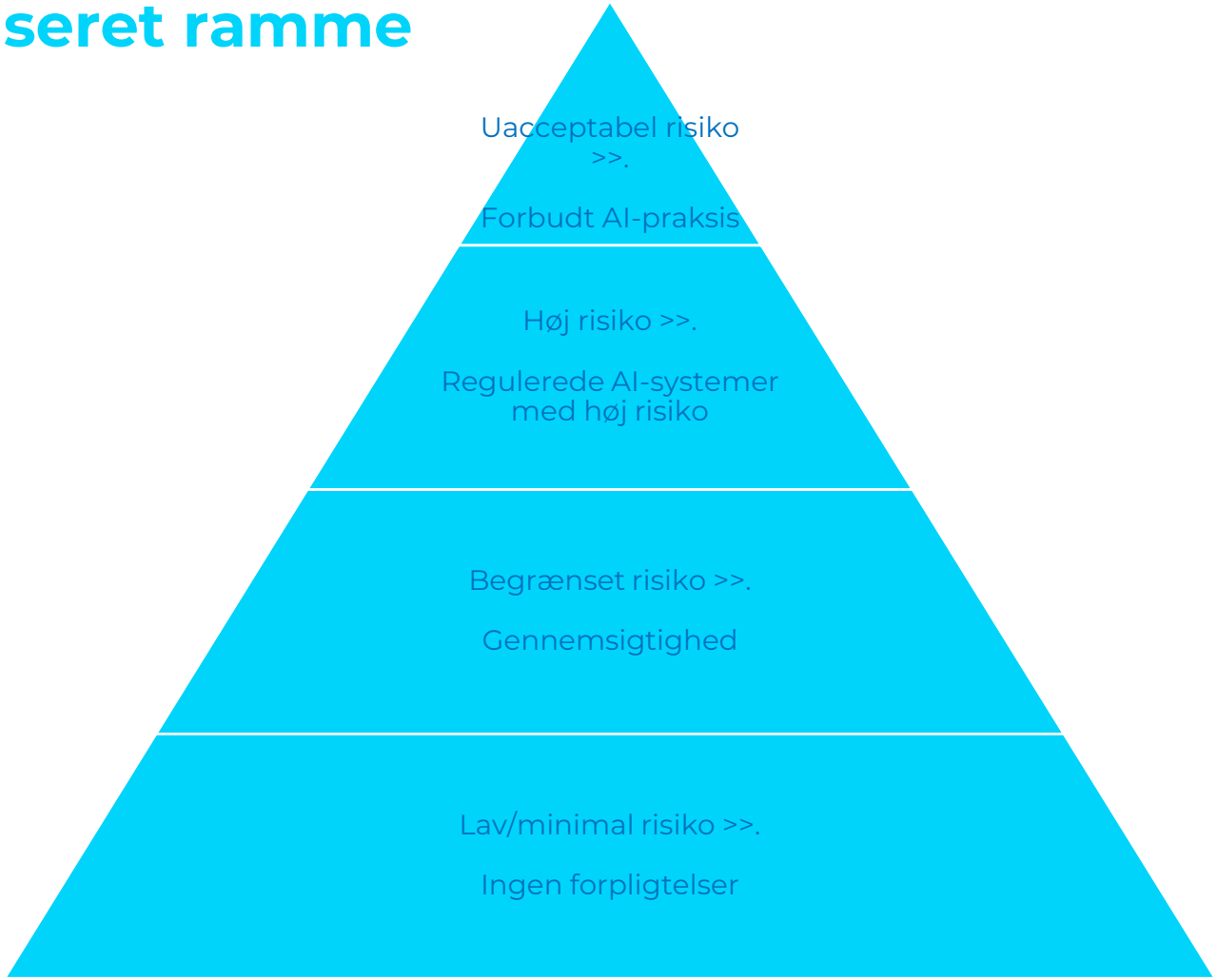
https://artificialintelligenceact.com
https://www.euaiact.com

EU’s forordning om kunstig intelligens

Risikobaseret ramme

De fastlagte etiske principper skal omsættes til endelige **krav** til realiseringen af troværdig AI. Disse krav gælder for alle deltagere i AI-systemets livscyklus:

- **Udviklere**: personer eller teams, der udfører AI-forskning, -design eller -udvikling.
- **Udbydere:** organisationer, der implementerer AI og tilbyder AI-drevne produkter eller services.
- **Slutbrugere:** Mennesker, der interagerer med AI-systemer, enten direkte eller indirekte.
- **Det bredere samfund:** dem, der påvirkes af AI, både direkte og indirekte.



EU's AI Act 'Pyramid of Risks' (Kilde: Europa-Parlamentet)

EU’s forordning om kunstig intelligens

Risikobaseret ramme | Minimal eller ingen risiko

AI-applikationer, der anses for at udgøre **en minimal eller ingen risiko** for borgernes rettigheder eller sikkerhed. Langt de fleste AI-systemer falder ind under kategorien minimal risiko.

Eksempler på AI-applikationer med minimal risiko:

- AI-aktiverede anbefalingssystemer
- Spam-filtre

Under de nuværende regler er disse systemer ikke underlagt nogen specifikke nye forpligtelser, men skal overholde allerede eksisterende love.

EU’s forordning om kunstig intelligens

Risikobaseret ramme | Begrænset risiko

AI-systemer **med begrænset risiko** kræver gennemsigtighed.

For disse AI-systemer gælder kravene om gennemsigtighed: Brugerne skal informeres om, at de interagerer med AI, og have de nødvendige oplysninger til at beslutte sig for fortsat brug.

Eksempler på AI-applikationer med begrænset risiko:

- AI, der genererer eller manipulerer billed- eller lydindhold.
- AI, der skaber eller ændrer videoindhold, f.eks. deepfakes.

EU's forordning om kunstig intelligens

Risikobaseret ramme | Høj risiko

AI-systemer med høj risiko kan have en betydelig indvirkning på en brugers liv.

Der er strenge krav til disse systemer, før de kan tages i brug i EU, herunder forpligtelser til risikostyring og datastyring.

Eksempler på AI med høj risiko omfatter:

- Styring og drift af kritisk infrastruktur
- Uddannelse og erhvervsuddannelse/lærepladser
- Beskæftigelse, arbejdsledelse og adgang til at drive selvstændig virksomhed
- Adgang til og brug af vigtige private services og offentlige services og ydelser
- Retshåndhævelse
- Forvaltning af migration, asyl og grænsekontrol
- Hjælp til juridisk fortolkning og anvendelse af loven.

EU's forordning om kunstig intelligens

Risikobaseret ramme | Uacceptabel risiko

AI-systemer med **uacceptabel risiko** anses for at udgøre en klar trussel mod menneskers grundlæggende rettigheder og er forbudt. Systemer med uacceptabel risiko er dem, der er udnyttende, manipulerende eller bruger subliminale teknikker.

Disse AI-systemer er forbudt at sælge på EU-markedet.

Eksempler på forbudt brug af AI inkluderer:

- Kognitiv adfærdsmæssig manipulation af mennesker eller specifikke sårbare grupper: for eksempel stemmeaktiveret legetøj, der opfordrer børn til farlig adfærd.
- Inddeling i sociale klasser: Klassificering af mennesker baseret på adfærd, socioøkonomisk status eller personlige egenskaber.
- Biometrisk identifikation og kategorisering af mennesker
- Biometriske identifikationssystemer i realtid og på afstand, f.eks. ansigtsgenkendelse

LIVSCYKLUS FOR AI-UDVIKLING

BILLEDKILDE | Freepik

Livscyklus for AI-udvikling

Omsætning af etiske principper, rammer og retningslinjer til konkrete handlinger

Betydningen af principper, rammer og retningslinjer for AI-udvikling er velkendt, men der er flere **forhindringer** for at anvende dem i praksis, når man skaber AI-løsninger:

- AI-principper har en tendens til at være abstrakte, forskelligartede og indviklede.
- Det er komplekst at omsætte nuanceret menneskelig adfærd til enkle, universelle værktøjer.
- Principper er ofte genstand for personlig fortolkning, som kan være påvirket af menneskers individuelle baggrunde.
- Der er et fravær af universel standardisering i regler, rammer og retningslinjer (selv om der er foretaget nogle generaliseringer som præsenteret tidligere).
- Selvom rammerne er nyttige til at identificere problemer, giver de ikke nødvendigvis handlingsorienteret vejledning om de specifikke skridt, der skal tages.

Morley (2019) / Prem (2023)

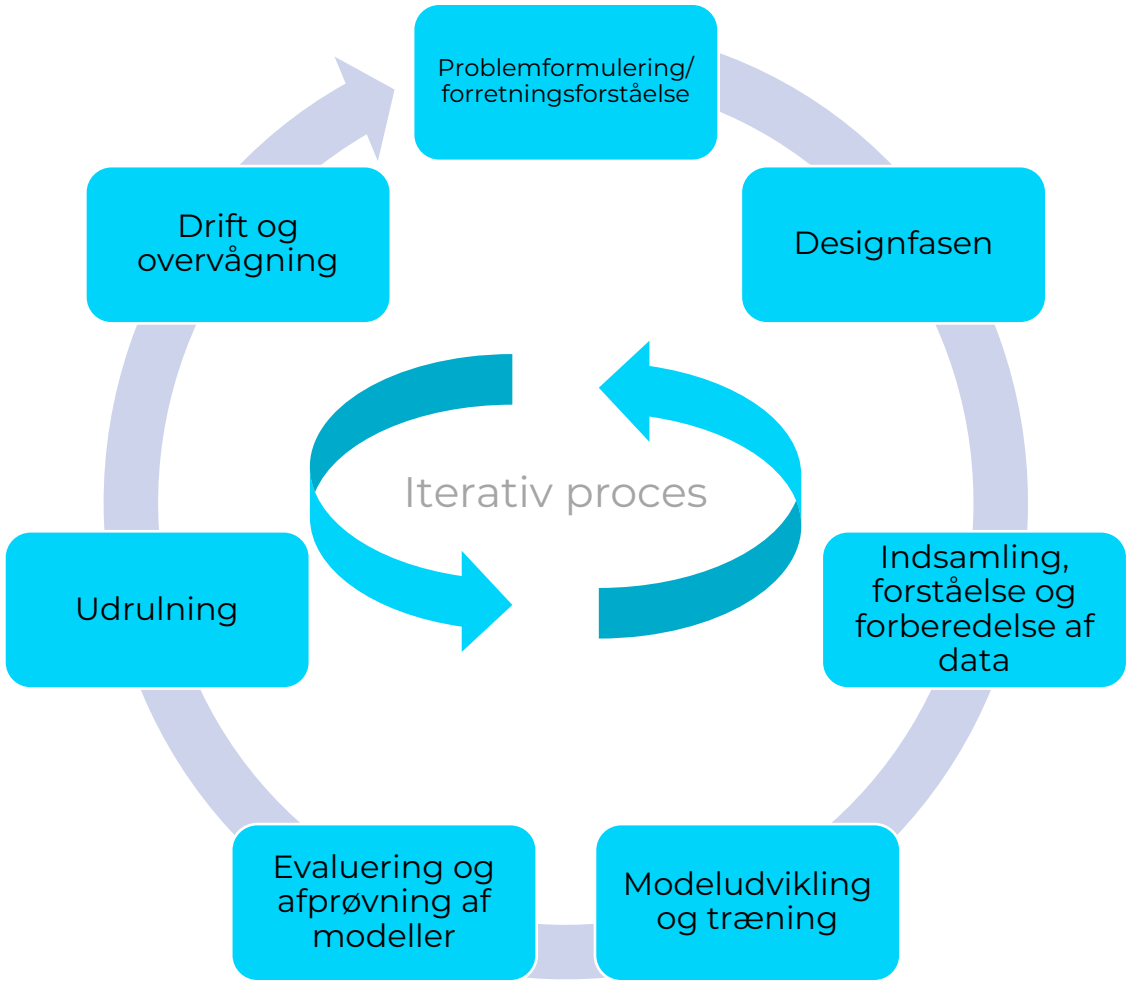
Livscyklus for AI-udvikling

Omsætning af etiske principper, rammer og retningslinjer til konkrete handlinger: Livscyklus for AI-udvikling

- Der er adgang til værktøjer, der er designet til at tackle etiske udfordringer i AI-løsninger, men de dækker ofte kun et udvalgt antal principper eller udviklingsstadier for AI-løsninger.
- For at omsætte etiske principper til konkrete handlinger anbefales det at se AI-udviklingsprocessen som en række trin, der starter med en business case og slutter med, at løsningen er i drift (livscyklus for AI-udvikling).
- Livscyklus for AI-udvikling anlægger et holistisk syn på etiske overvejelser gennem alle udviklingsstadier

Morley (2019) / Prem (2023)

Livscyklus for AI-udvikling



- Som i ethvert andet projekt anbefales det, at AI-designprojekter følger designtrin for at omdanne innovative ideer til praktiske og veldefinerede løsninger.
- Det er vigtigt at forstå, at hvert trin spiller en afgørende rolle i udviklingen af løsningen og danner en AI-udviklingslivscyklus.
- Processen er ITERATIV, hvilket betyder, at der kan være behov for at gå tilbage til det forrige trin, hvis det er nødvendigt, og ændre løsningen.

De designfaser for AI-udvikling, der præsenteres her, er en kombination af de forskellige modeller, der er beskrevet i artiklerne Data Science Process Alliance (2023), Morley et al. (2019), Prehm's (2022) og Rochel et al. (2020).

Data Science Process Alliance /Morley (2019) / Prem (2023) / Rochel & Evequoz (2020)

Livscyklus for AI-udvikling

PROBLEMFOR-MULERING OG FORRETNINGS-FORSTÅELSE

En definition af problemet er afgørende for at forstå kundernes behov og AI-løsningens rolle.

DESIGN-FASE

Designfasen omdanner forretningskravene til designkrav for udviklerne
.

INDSAMLING, FORSTÅELSE OG FORBEREDELSE AF DATA

Dataindsamling, -forståelse og -forberedelse er grundlaget for løsningen

MODELUDVIKLING OG TRÆNING

Modeludvikling og -evaluering har til formål at omdanne data til en gennemførlig løsning på det oprindelige problem.

EVALUERING OG AFPRØVNING AF MODELLER

Modellen skal testes grundigt, og funktionaliteten skal afspejles i kundernes behov

UDRULNING

Implementering er at bringe løsningen ud til brugerne

DRIFT OG OVERVÅGNING

Løsningen skal overvåges over tid og vedligeholdes i henhold til skiftende behov.

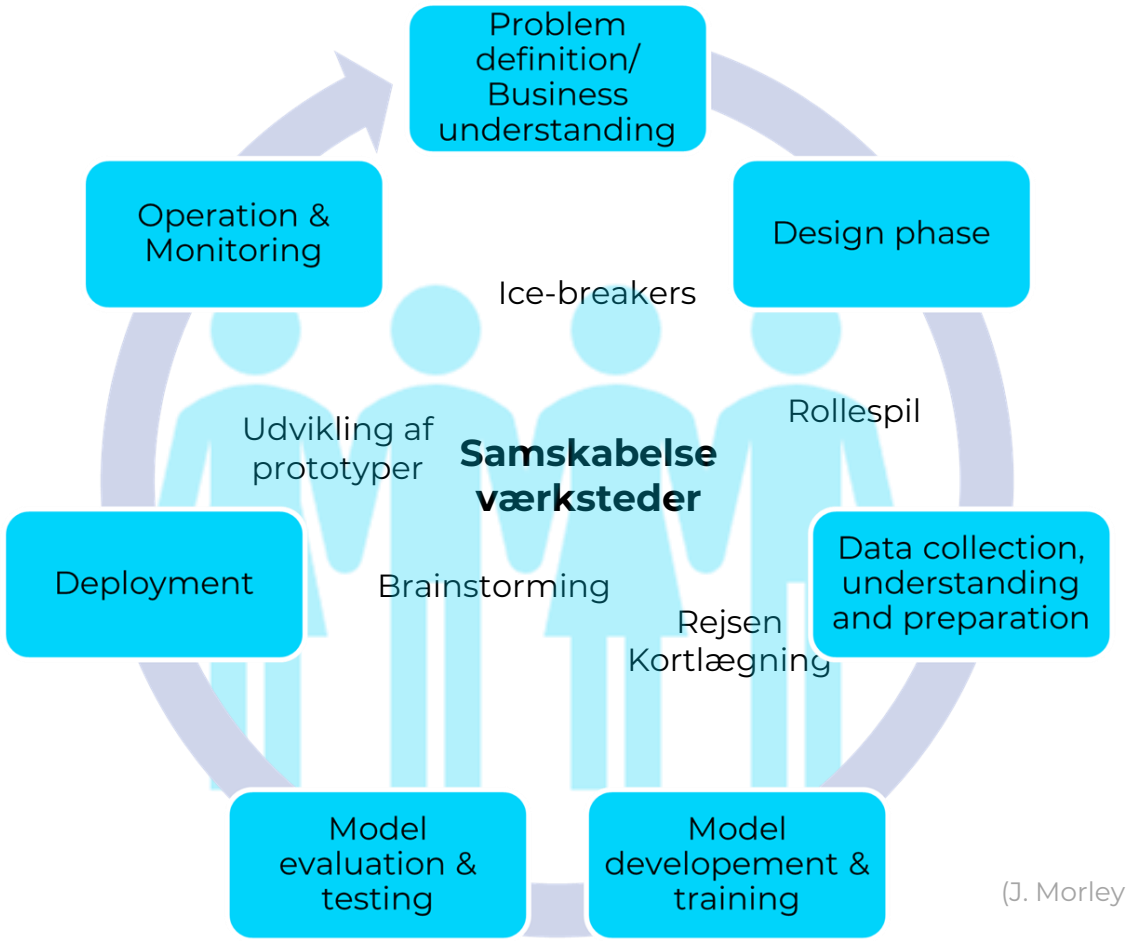
Data Science Process Alliance /Morley (2019) / Prem (2023) / Rochel & Evequoz (2020)

INDDRAGELSE AF INTERESSENTER I UDVIKLINGEN AF AI-LØSNINGER

BILLEDKILDE | Freepik

Inddragelse af interessenter i udviklingen af AI-løsninger

Hvorfor er det vigtigt at engagere forskellige interessenter i udviklingen af AI-løsninger?

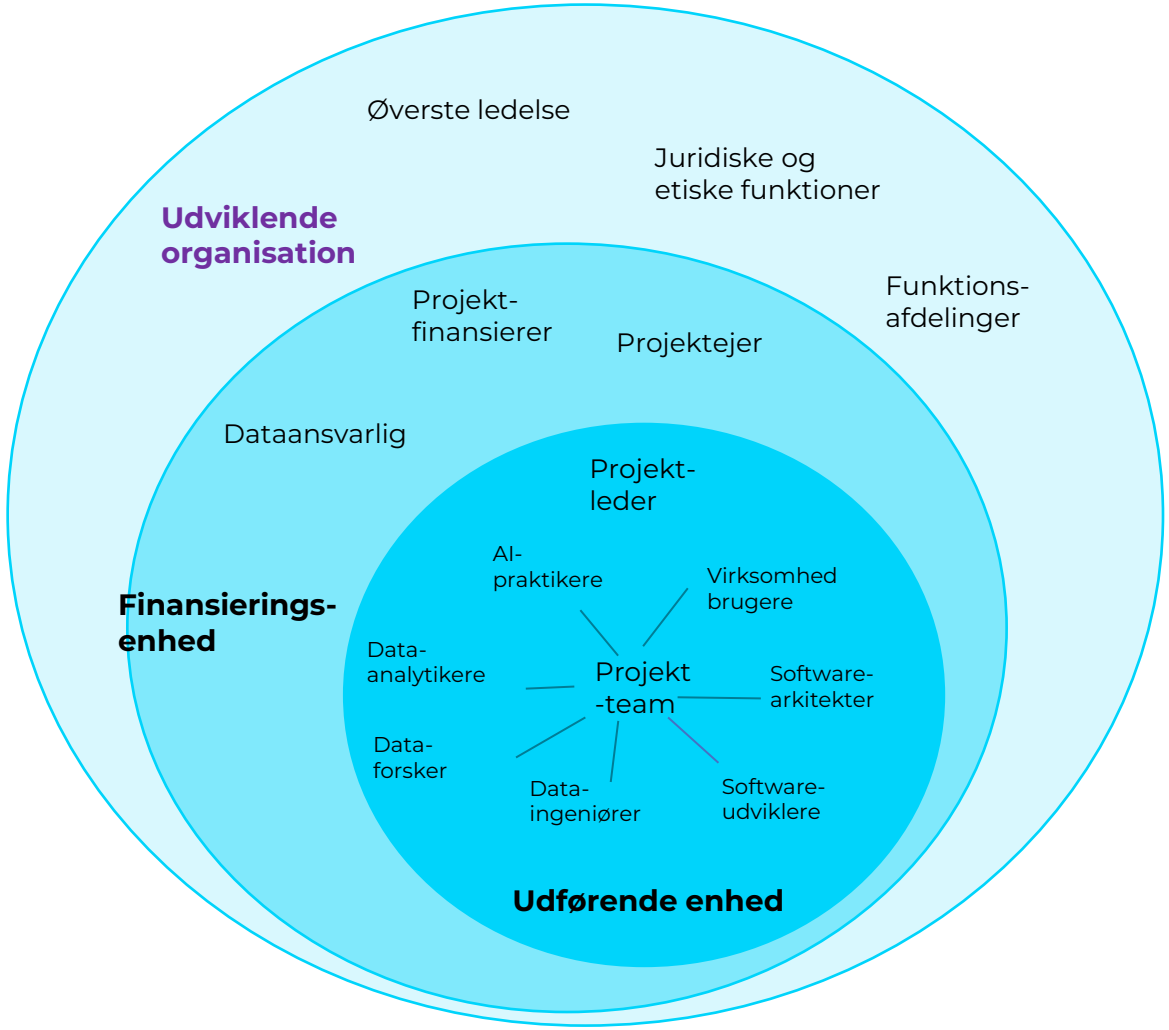


- For at sikre, at konsekvenserne af AI-udvikling overvejes fra forskellige synsvinkler, og for at udvikle pålidelige systemer, er det vigtigt at inddrage og konsultere flere interessenter i udviklingsprocessen, herunder dem, der vil blive påvirket af AI-løsningen.
- Samskabelse er en praksis, der bruges til at samarbejde med interessenter, herunder slutbrugere eller kunder, under en designproces for at dele viden mellem interessenter med forskellige roller og ekspertise og gør det muligt bedre at forstå slutbrugernes behov og dermed designe løsninger, der tager hensyn til forskellige brugere.

(J. Morley et al., 2019) / (UNESCO, 2023) / (Interaction Design Foundation, Co-creation) / (Anika Sanin, 2020)

Inddragelse af interessenter i udviklingen af AI-løsninger

Interessenter inden for udvikling



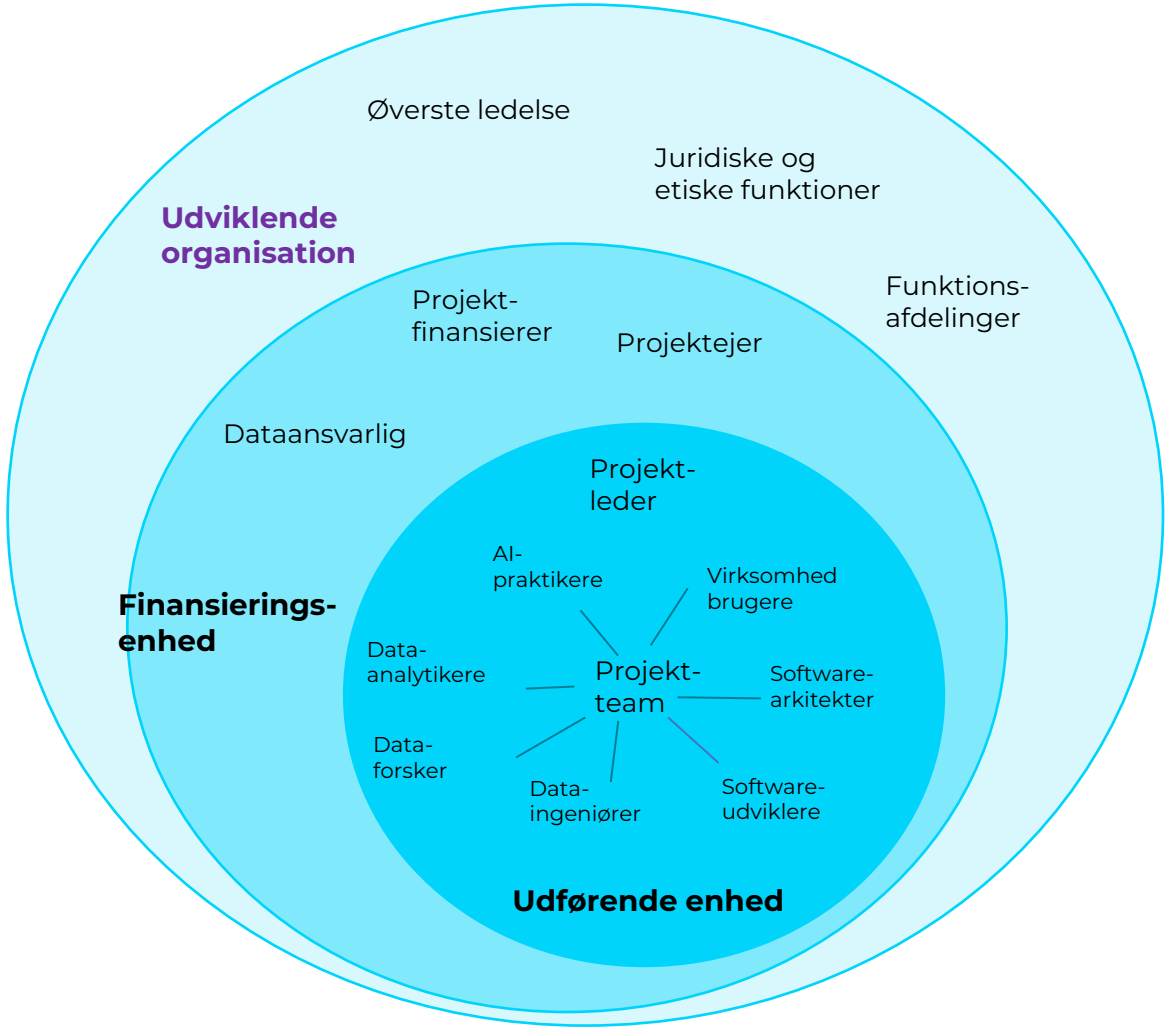
Udviklende organisation

- Støtter AI-projektet økonomisk og fastlægger projektets omfang.
- Projektroller er den **øverste ledelse, funktions-afdelinger og juridiske og etiske funktioner:**
  - **Ledelsen** er ansvarlig for virksomhedens strategiske, finansielle og udviklingsmæssige beslutninger.
  - En **funktionsafdeling** er en specialiseret organisatorisk enhed i en virksomhed, der fokuserer på et specifikle opgaver eller aktiviteter, ofte grupperet ud fra fælles funktioner som f.eks. økonomi, marketing eller HR.
  - **Juridiske og etiske funktioner** i en organisation involverer forskellige roller og ansvarsområder for at sikre overholdelse af love, regler og etiske standarder.

(G.J. Miller, 2022)

Inddragelse af interessenter i udviklingen af AI-løsninger

Interessenter inden for udvikling



Finansieringsenhed

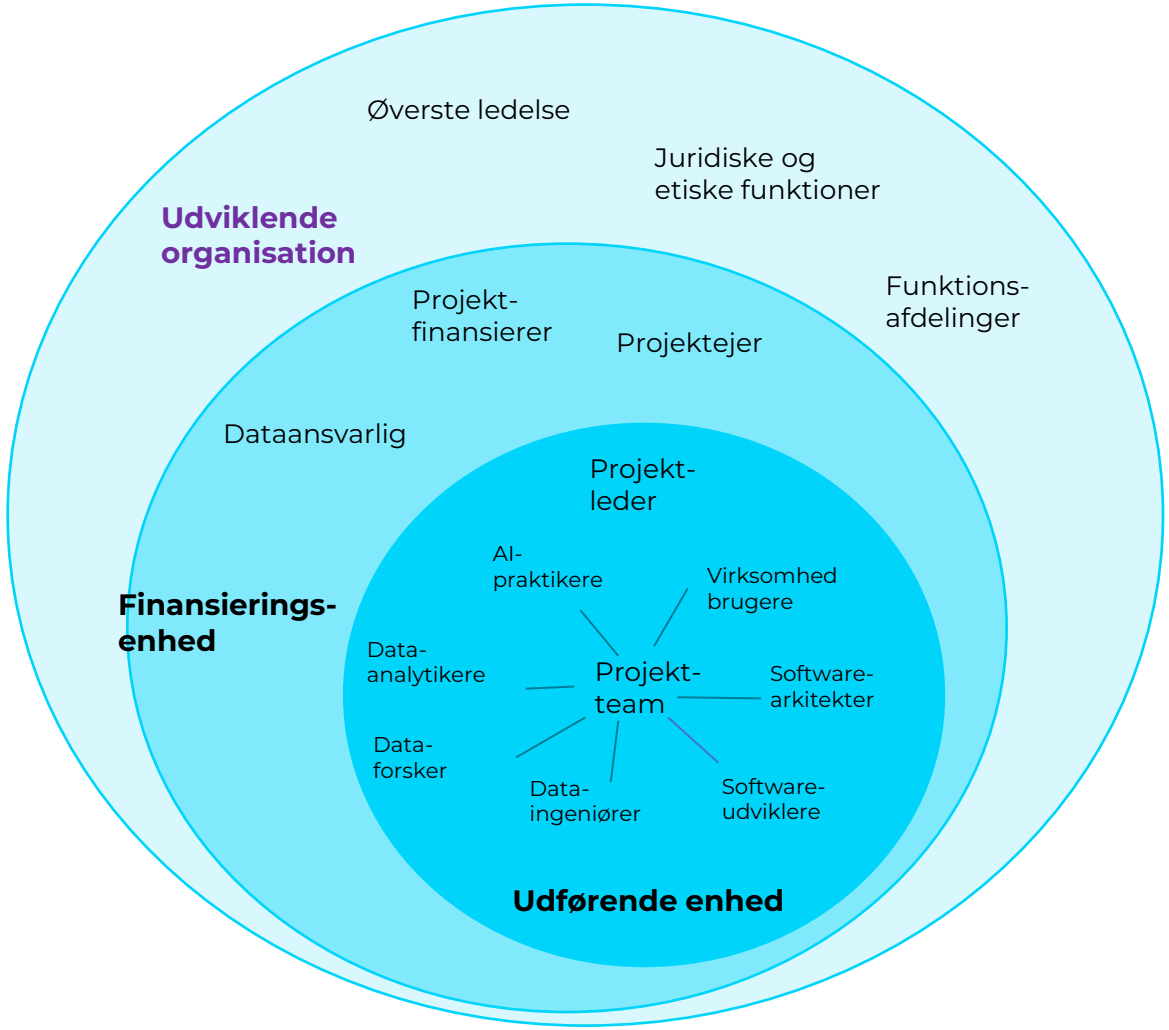
Projektrollerne er **projektfinansierer**, **projektejer og dataansvarlig**.

- **Projektfinansiereren** finansierer projektet og tildeler ressourcer til projektet.
- **Projektejer** tilbyder strategisk retning.
- **Den dataansvarlige** er ansvarlig for forvaltningen af dataene og deres brug i systemudviklingen.

(G.J. Miller, 2022)

Inddragelse af interessenter i udviklingen af AI-løsninger

Interessenter i udvikling



Udførende enhed

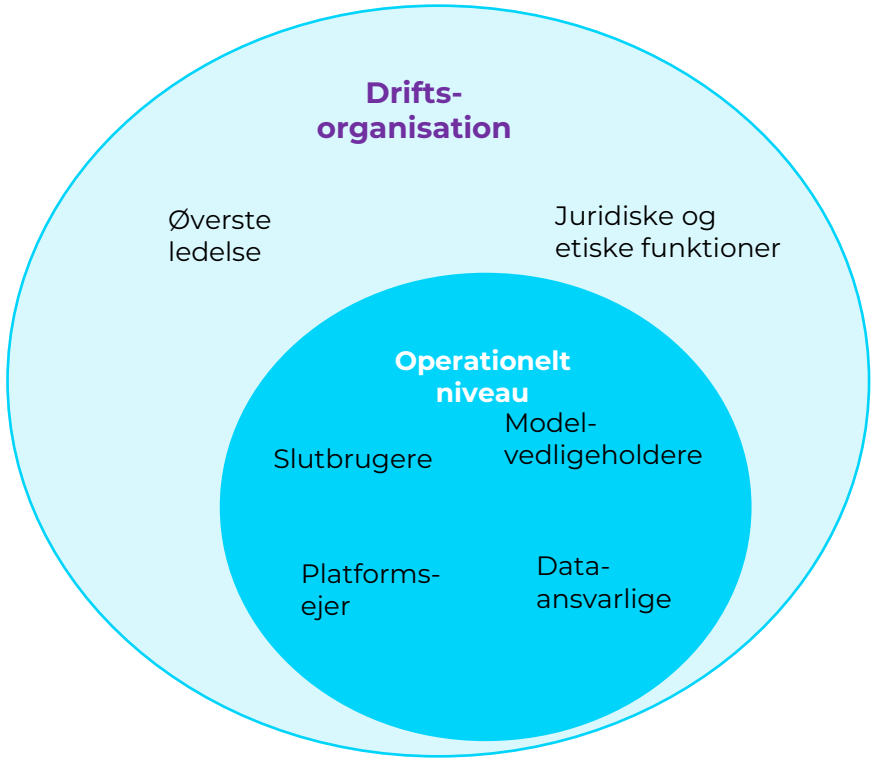
Projektrollerne er **projektledere** og **projektteam**.

- **Projektlederen** er ansvarlig for projektets resultater og sikrer, at etiske, privatlivs- og sikkerhedsmæssige standarder og forventninger respekteres, og at relevante interessenter inddrages.
- **AI-projektteamet** består af softwarearkitekter, softwareudviklere, dataingeniører, dataforskere, dataanalytikere, AI-praktikere og slutbrugere.

(G.J. Miller, 2022)

Inddragelse af interessenter i udviklingen af AI-løsninger

Brug af interessenter



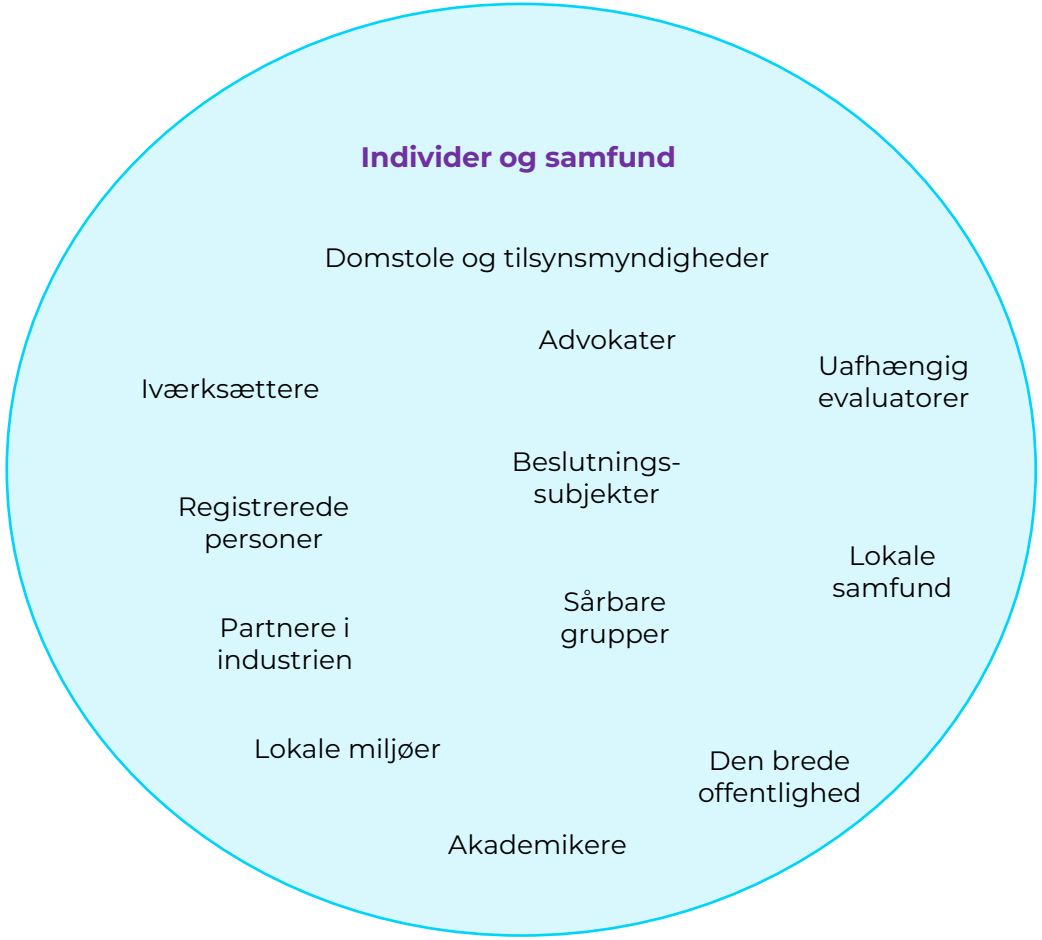
Driftsorganisation

- En driftsorganisation er en kundeinteressent.
- Ligesom i udviklingsorganisationen omfatter driftsorganisationens projektroller den **øverste ledelse samt juridiske og etiske funktioner.**
- På driftsniveau er projektrollerne **slutbrugere, modelvedligeholdere**, **dataansvarlige** og **platformsejere**.
    - **Slutbrugere** er enkeltpersoner eller grupper, der interagerer med AI-systemet gennem arbejds- eller servicekontrakter med driftsorganisationen eller som forbrugere.
    - **Modelvedligeholdere** er ansvarlige for den løbende administration, opdatering og optimering af de maskinlæringsmodeller, der bruges i AI-systemet.
    - **Dataansvarlige** bestemmer formålene med og midlerne til behandling af personoplysninger inden for AI-systemet.
    - **Platformsejere** er ansvarlige for den overordnede infrastruktur og implementeringen af AI-systemet.

(G.J. Miller, 2022)

Inddragelse af interessenter i udviklingen af AI-løsninger

Eksterne interessenter



Enkeltpersoner og samfund

- **Enkeltpersoner** er mennesker, der har aftaler med udviklings- eller driftsorganisationen, f.eks. via service- eller ansættelseskontrakter. Denne gruppe omfatter datasubjekter, beslutningssubjekter og medarbejdere, og de betragtes som passive interessenter.
- **Samfundet** består af enkeltpersoner og offentligheden. De kan blive negativt påvirket af udviklingen eller brugen af AI-systemet**.**
- Denne gruppe omfatter lokalsamfund, miljøer og sårbare grupper som mennesker med handicap, minoritetsgrupper, mindreårige og den brede offentlighed.

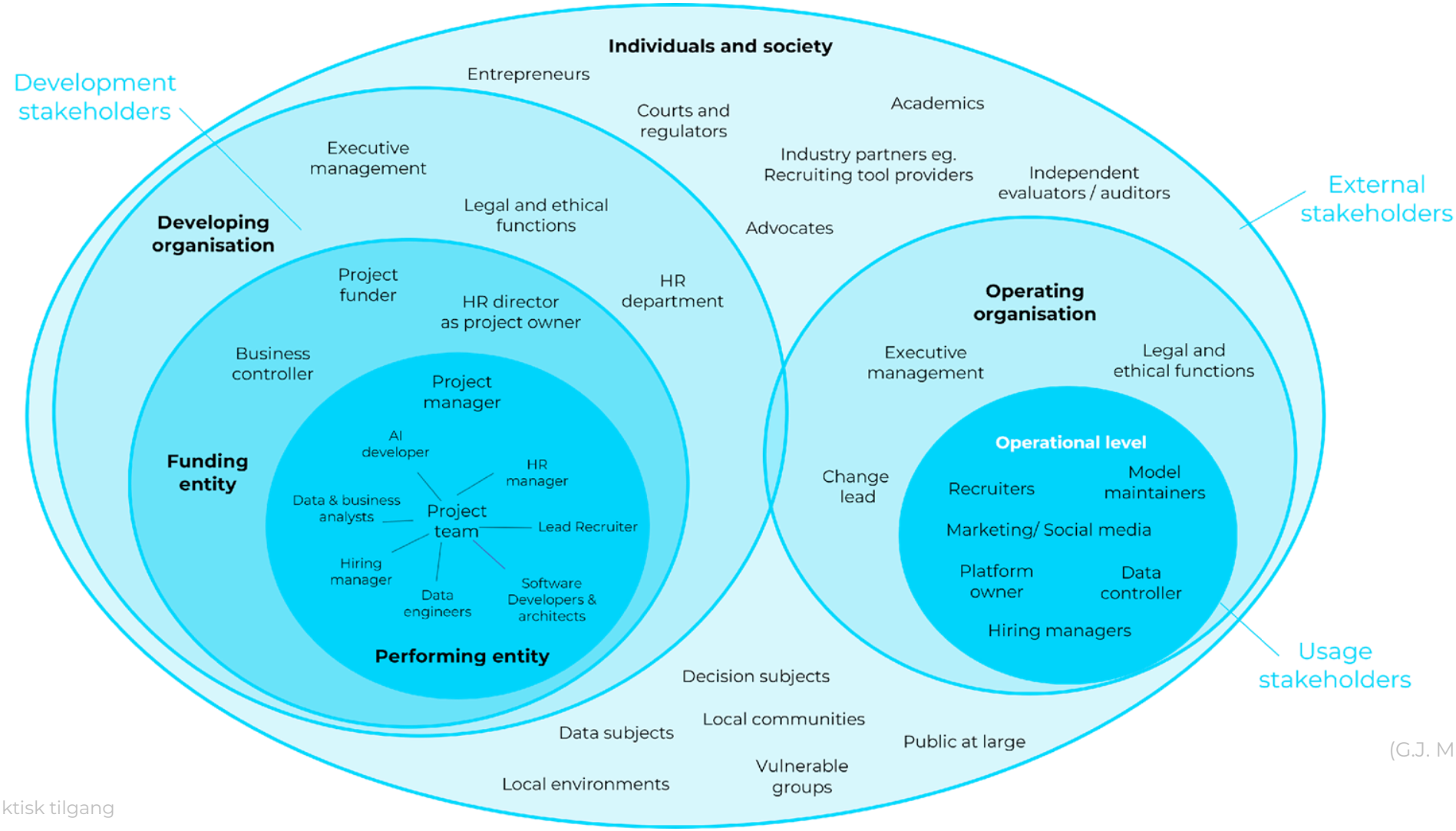
AI-ingeniører er en del af en større samfundsmæssig diskussion, hvor de skal spille en aktiv rolle i at integrere andre perspektiver.

- **Repræsentanter** er grupper og organisationer, som ikke er berørt af AI-systemet, men som kan repræsentere andre i projektet.
- **Formelle repræsentanter** omfatter domstole, love, regulatorer og fagforeninger.
- **Iværksættere** kan være involveret i udviklingen af AI-systemet.
- **Uafhængige evaluatorer** er personer eller grupper udefra, f.eks. journalister, revisorer eller anmeldere.

(G.J. Miller, 2022), (Rodrigues, 2020)

Inddragelse af interessenter i udviklingen af AI-løsninger

Interessenter i AI-udvikling

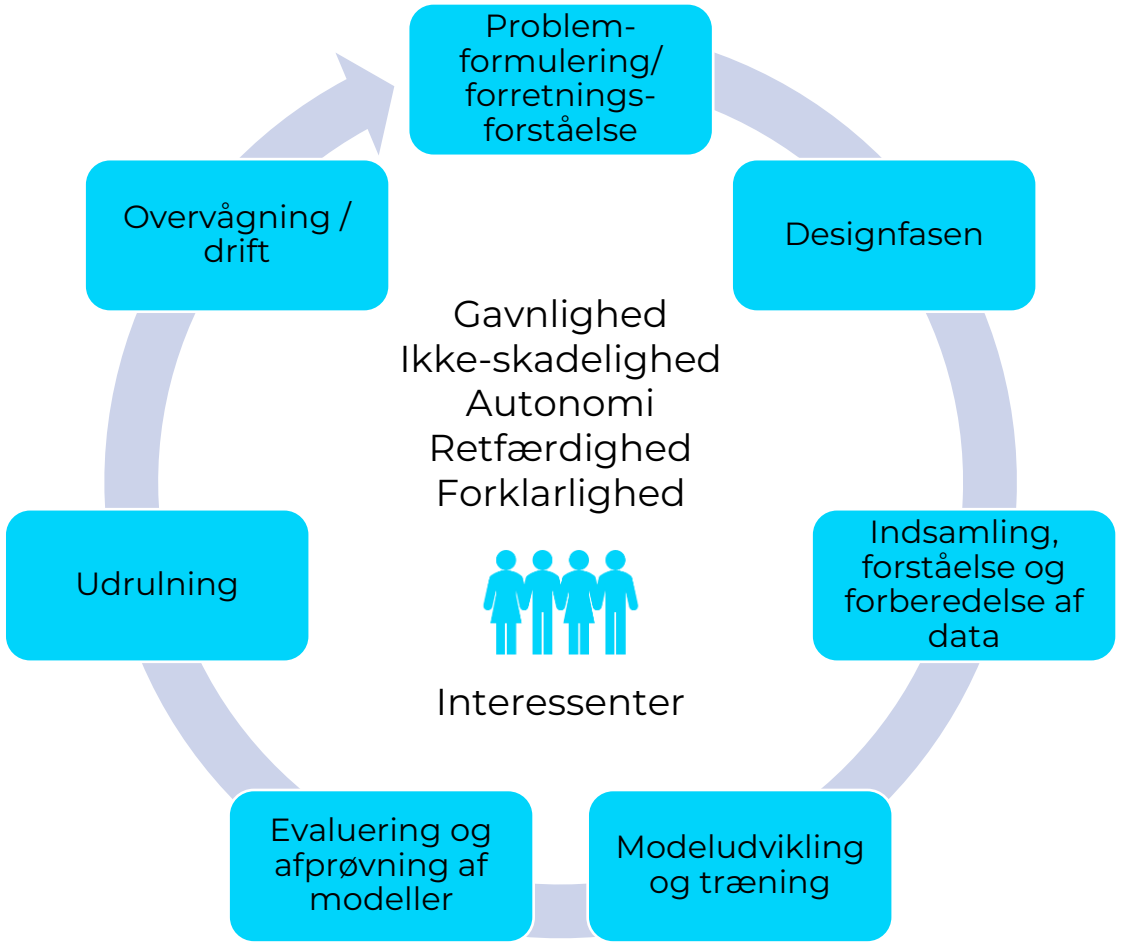


(G.J. Miller, 2022)

ETISKE PRINCIPPER OG INDDRAGELSE AF INTERESSENTER I I AI-UDVIKLINGENS LIVSCYKLUS

BILLEDKILDE | Freepik

Etiske principper og interessenter i AI-udviklingens livscyklus



Gennem hele AI-livscyklussen er det vigtigt at holde fokus på etiske overvejelser og inddragelse af interessenter.

På de følgende sider gennemgås livscyklussen fra tre perspektiver:

1. **Beskrivelse af situationen**
2. **De vigtigste etiske principper**
3. **Involverede interessenter**

Ikke alle etiske principper har samme vægt i alle sammenhænge eller på alle stadier af udviklingen, og derfor er det kun de vigtigste overvejelser og synspunkter, der præsenteres i det følgende.

Data Science Process Alliance /Morley (2019) / Prem (2023) / Rochel & Evequoz (2020)

Etiske principper og interessenter i AI-udviklingens livscyklus

Problemformulering og forretningsforståelse

Den afgørende første fase for et vellykket projekt

- Forstå det problem, der skal løses, eller den mulighed, der skal undersøges.
- Gennemfør research og interviews med interessenter for at formulere en klar problemstilling.
- Forstå kundernes behov, og hvordan de omsættes til en levedygtig forretningsløsning.
- Definér mål og succeskriterier for løsningen.
- Gennemfør en forundersøgelse for at afgøre, om AI er den mest hensigtsmæssige løsning.
- Vurder både den tekniske og etiske gennemførlighed af projektet

Data Science Process Alliance /Morley (2019) / DeSilva & Alahakoon (2022)

Etiske principper og interessenter i AI-udviklingens livscyklus

Problemformulering og forretningsforståelse

Etiske principper, der skal overvejes

PROBLEM-FORMULERING OG FORRETNINGS-FORSTÅELSE	DESIGN-FASE	DATA-INDSAMLING, DATAFORSTÅELSE OG DATA-FORBEREDELSE	MODELUDVIKLING OG TRÆNING	EVALUERING OG AFPRØVNING AF MODELLER	UDRULNING	DRIFT OG OVERVÅGNING

At lægge fundamentet - tage fat på grundlæggende etiske spørgsmål som menneskerettigheder, miljøpåvirkning og samfundsmæssige hensyn.

Gavnlighed - sørg for, at AI-løsningen forbedrer den enkeltes og samfundets bedste. Overvej følgende områder:

- **Deltagelse af interessenter**
  - Opbyg troværdige systemer, der understøtter menneskelig trivsel.
  - Inddrag de berørte personer i udviklingsprocessen.
- **Beskyttelse af grundlæggende** rettigheder
  - Sikr individuelle rettigheder og styrk den etiske integritet.
- **Miljømæssig** bæredygtighed
  - Minimer miljøpåvirkningen gennem en bæredygtig praksis.
- **Begrundelse**
  - Definer klart og kæd systemets formål sammen med konkrete samfundsmæssige fordele.

Morley (2019) / Prem (2023)

Etiske principper og interessenter i AI-udviklingens livscyklus

Problemformulering og forretningsforståelse

Etiske principper, der skal overvejes

PROBLEM-FORMULERING OG FORRETNINGS-FORSTÅELSE	DESIGN-FASE	DATA-INDSAMLING, DATAFORSTÅELSE OG DATA FORBEREDELSE	MODELUDVIKLING OG TRÆNING	EVALUERING OG AFPRØVNING AF MODELLER	UDRULNING	DRIFT OG OVERVÅGNING

Ikke-skadelighed

- Sørg for, at problemformuleringen og den foreslåede løsning ikke skader enkeltpersoner eller samfund.
- Overvej og overvåg nøje systemets indvirkning på det fysiske og mentale velbefindende.
- Vurder potentielle sikkerhedsrisici og strategier for afhjælpning.

Retfærdighed

- Evaluer systemets indvirkning på institutioner, demokrati og samfund.
- Overvej, om visse grupper favoriseres eller forfordeles.
- Undersøg, om projektet kan krænke den individuelle autonomi.

Morley (2019)

Etiske principper og interessenter i AI-udviklingens livscyklus

Problemformulering og forretningsforståelse

Interessenter

- **Slutbrugere:** primære brugere af AI-løsningen
- **AI/datavidenskabsfolk**: formulerer problemet og konceptualiserer en løsning
- **De berørte samfundsgrupper**
- **Domæneeksperter:** dyb viden om den branche, hvor AI anvendes
- **AI-udviklere**: tidlig involvering giver dem en dyb forståelse af problemet
- **Projektledere:** fører tilsyn med hele projektets forløb
- **Forretningsinteressenter:** sikrer overensstemmelse med forretningsmål og -ressourcer
- **Etikere og samfundsforskere**: Fremhæver potentielle konsekvenser

(D. De Silva & D. Alahakoon, 2022)

Etiske principper og interessenter i AI-udviklingens livscyklus

Designfasen

Denne fase omdanner det grundlæggende arbejde, der blev lagt i problemformuleringsfasen, til detaljerede planer for opbygning og implementering af AI-systemet.

- Specificer, hvad projektet skal munde ud i (mål og resultater)
- Vurder de økonomiske omkostninger i forhold til de forventede fordele.
- Skitsér de funktioner, som AI-systemet skal have.
- Beslut dig for tilgangen til, hvordan data udnyttes, f.eks. brug af prætrænede modeller.
- Etabler metrikker til at måle projektets succes.
- Integrer etiske retningslinjer i udviklingsprocessen.
- Lav en tidsplan og styr inddragelsen af interessenter.

Rochel (2020)/ Morley (2019) / & DeSilva&Alahakoon (2022)

Etiske principper og interessenter i AI-udviklingens livscyklus

Designfasen

Etiske principper, der skal overvejes

PROBLEM-FORMULERING OG FORRETNINGS-FORSTÅELSE	DESIGN-FASE	DATA-INDSAMLING, DATAFORSTÅELSE OG DATA FORBEREDELSE	MODELUDVIKLING OG TRÆNING	EVALUERING OG AFPRØVNING AF MODELLER	UDRULNING	DRIFT OG OVERVÅGNING

Retfærdighed | Undgå uretfærdige bias

- **Krav til design**
  - Implementer mekanismer til at identificere og afbøde bias i data og algoritmer.
  - Tilstræb retfærdige resultater på tværs af forskellige grupper.
- **Bevidsthed hos udviklere**
  Sikr, at udviklere er bevidste om potentielle bias, især i forbindelse med f.eks. race, hudfarve, slægt, køn, alder, sprog, religion, politisk overbevisning, national, etnisk og social oprindelse, økonomisk status, fødselsforhold og handicap.
- **Kommunikation og respons**
  Etabler klare retningslinjer for kommunikation, indrapportering af og reaktion på bias, negative påvirkninger eller nødvendige kompromiser.
- **Håndtering af data**
  Overvej datalagringssystemer for at forhindre siloopdelte data.
- **Risikostyring**
  Udvikl "Hvad sker der, hvis"-foranstaltninger til at håndtere uventede resultater.

Morley (2019) / UNESCO (2023)

Etiske principper og interessenter i AI-udviklingens livscyklus

Designfasen

Etiske principper, der skal overvejes

PROBLEM-FORMULERING OG FORRETNINGS-FORSTÅELSE	DESIGN-FASE	DATA-INDSAMLING, DATAFORSTÅELSE OG DATA FORBEREDELSE	MODELUDVIKLING OG TRÆNING	EVALUERING OG AFPRØVNING AF MODELLER	UDRULNING	DRIFT OG OVERVÅGNING

Forklarlighed/gennemsigtighed
- Sørg for klar dokumentation og brugervenlige forklaringer af AI-beslutninger.
- Sikr gennemsigtighed i datakilder og algoritmiske processer.

Gavnlighed
Inddrag interessenter i udviklingsprocessen for at overveje forskellige etiske aspekter.

Autonomi
Implementer mekanismer for menneskeligt tilsyn for at bevare menneskelige værdier i AI-løsninger.

HLEG/ Unesco / Morley / Prem

Etiske principper og interessenter i AI-udviklingens livscyklus

Designfasen

Interessenter

- **Softwarearkitekter:** lægger det tekniske fundament for AI-løsningen og tager højde for både nuværende og fremtidige behov.
- **UX-designere (User Experience):** afgørende for at skabe grænseflader, der øger brugernes accept og tilfredshed, og for at sikre, at brugerne kan interagere effektivt med AI-systemet.
- **Etiske eksperter:** adresserer etiske aspekter.
- **AI-praktikere:** samarbejder med UX-designere om at integrere AI-komponenter problemfrit.
- **Juridiske og etiske funktioner:** mindsker juridiske risici ved at adressere compliance-krav tidligt i designfasen og hjælper med at designe løsninger, der respekterer databeskyttelses- og privatlivsbestemmelser.
- **Mennesker og lokalsamfund, der er berørt af AI-modellen**: participatorisk designtilgang.

(D. De Silva & D. Alahakoon, 2022) / (Peter, 2024) / (White, 2024)

Etiske principper og interessenter i AI-udviklingens livscyklus

Indsamling, forståelse og forberedelse af data

Når målet og planen for at nå det er fastlagt, er næste skridt at arbejde med relevante data.

- Indsaml alle nødvendige data til projektet.
- Beskriv, udforsk, verificer, udvælg, rens, konstruer, integrer og formater data i passende kategorier og koncepter til det tilsigtede formål.
- Håndter og løs problemer i forbindelse med manglende data for at sikre fuldstændighed.
- Anerkend og forstå alle antagelser, der ligger til grund for de eksisterende data.
- Afklar, hvordan datasættene vil blive brugt til beslutningstagning.
- Opretholdt en høj datakvalitet, især i tilfælde af uklare eller manglende oplysninger.

Dette er en tidskrævende fase, hvor fundamentet for den fungerende løsning bygges. Løsningens pålidelighed og ydeevne hænger direkte sammen med kvaliteten af de underliggende data.

Etiske principper og interessenter i AI-udviklingens livscyklus

Indsamling, forståelse og forberedelse af data

Etiske principper, der skal overvejes

PROBLEM-FORMULERING OG FORRETNINGS-FORSTÅELSE	DESIGN-FASE	DATA-INDSAMLING, DATAFORSTÅELSE OG DATA FORBEREDELSE	MODELUDVIKLING OG TRÆNING	EVALUERING OG AFPRØVNING AF MODELLER	UDRULNING	DRIFT OG OVERVÅGNING

Ikke-skadelighed (*non-maleficence*) 1/2

- **Dataintegritet og -kvalitet som kan forårsage bias**
  - Håndter potentielle bias, der favoriserer visse demografiske grupper. Sørg for at være nøjagtig, når du kombinerer datasæt (f.eks. bør personer med samme navn ikke blandes).
  - Udfør grundig *due diligence* for at vurdere og verificere dataenes nøjagtighed og sikre, at de ikke opretholder uretfærdige bias.
  - Undgå at bruge silodata baseret på specifik demografi, som kan forstærke bias.
  - Implementer et samlet datalager, f.eks. et datalager, for at minimere bias (bemærk: dyrt).

Kan bias være en relevant faktor i AI-løsningen? F.eks. påvirkes bilforsikringer af typer af ulykker og den involverede demografi. Juridiske begrænsninger: Køn kan ikke afgøre bilforsikringspriser, ifølge EU-Domstolen, 2012.

HLEG (2019) / De Silva & Alahakoon (2022) / Rochel et. Al (2023)

Etiske principper og interessenter i AI-udviklingens livscyklus

Indsamling, forståelse og forberedelse af data

Principper

PROBLEM-FORMULERING OG FORRETNINGS-FORSTÅELSE | DESIGN-FASE | **DATA-INDSAMLING, DATAFORSTÅELSE OG DATA FORBEREDELSE** | MODELUDVIKLING OG TRÆNING | EVALUERING OG AFPRØVNING AF MODELLER | UDRULNING | DRIFT OG OVERVÅGNING

Ikke-skadelighed (*non-maleficence*) 2/2

- **Databeskyttelse og fortrolighed**
  - Håndter typiske risici omkring databeskyttelse og fortrolighed under dataindsamlingen.
  - Sørg for, at AI-systemer opretholder datasikkerhed i hele livscyklussen og overholder regler som GDPR.

- **Modstandsdygtighed over for angreb og sikkerhed**
  - Beskyt datasæt mod angreb, og håndter potentielle sikkerhedsrisici for at opretholde en robust datasikkerhed.

HLEG (2019) / De Silva & Alahakoon (2022) / Rochel et al. (2023)

Etiske principper og interessenter i AI-udviklingens livscyklus

Indsamling, forståelse og forberedelse af data

Principper

PROBLEM-FORMULERING OG FORRETNINGS-FORSTÅELSE	DESIGN-FASE	DATA-INDSAMLING, DATAFORSTÅELSE OG DATA FORBEREDELSE	MODELUDVIKLING OG TRÆNING	EVALUERING OG AFPRØVNING AF MODELLER	UDRULNING	DRIFT OG OVERVÅGNING

- **Dataforskere** finder mønstre og tendenser i data for at omdanne dem til viden.
- **AI-ingeniører**: ansvarlige for at vurdere og verificere data.
- **Dataanalytikere** bidrager til at forstå dataenes egenskaber og hjælper med at informere om strategier for dataforberedelse.
- **Dataingeniører**: ansvarlige for at forberede dataene og sikre deres kvalitet.
- **Databeskyttelsesrådgiver (DPO):** Ansvar for at sikre, at deres organisation følger reglerne, når de håndterer persondata om medarbejdere, kunder, leverandører eller andre personer (kaldet registrerede).
- **Etikere:** tager stilling til etiske overvejelser i forbindelse med håndtering og udvælgelse af data.

(Miller, 2022), (J. Rochel & F. Evéquoz, 2020), (European Data Protection Supervisor, 2024)

Etiske principper og interessenter i AI-udviklingens livscyklus

Modeludvikling og træning

I denne fase udvikles en AI-model til at løse det definerede problem. Dette omfatter flere vigtige trin:

- Vælg og konfigurer den algoritmiske model og værktøjerne, og definer brugen af dem på forhånd.
- Angiv og vurder konsekvenserne af de valgte værktøjer, især med hensyn til potentielle ulemper og kompromiser.
- Beslut, hvilke elementer i modellen der skal styrkes, og hvilke der kan blive sekundære eller endda skjulte.
- Deltag i flere runder af videreudvikling under træningsprocessen, så modellen hele tiden forbedres.
- Valg af værktøjer indebærer ofte en afvejning af modstridende interesser. For eksempel skal man beslutte, om det er bedre at risikere, at spammails havner i den egentlige indbakke, eller at gyldige e-mails havner i spammappen.

Denne fase er afgørende, da den har direkte indflydelse på AI-løsningens effektivitet.

Rochel & Evéquoz (2022), Data Science Process Alliance

Etiske principper og interessenter i AI-udviklingens livscyklus

Modeludvikling og træning

Etiske principper, der skal overvejes

PROBLEM-FORMULERING OG FORRETNINGS-FORSTÅELSE | DESIGN-FASE | DATA-INDSAMLING, DATAFORSTÅELSE OG DATA FORBEREDELSE | **MODELUDVIKLING OG TRÆNING** | EVALUERING OG AFPRØVNING AF MODELLER | UDRULNING | DRIFT OG OVERVÅGNING

Forklarlighed

- Mens vi fokuserer på de tekniske aspekter, er det afgørende at tjekke forklarligheden, og om der findes bias i AI-systemet, hvilket giver en god mulighed for at gribe ind.
- Opretholdt gennemsigtighed i hele udviklingsprocessen for at sikre, at interessenterne kan forstå den. Det indebærer klar dokumentation af datakilder, modelvalg og rationalet bag disse beslutninger.

Retfærdighed

- Erkend og indrapporter eventuelle kompromiser mellem effektivitet og andre krav.
- Etabler en proces til at anerkende og evaluere effekten af disse kompromiser.
- AI-ingeniører skal begrunde deres beslutninger og tage højde for de potentielle negative konsekvenser for de berørte personer.

Morley (2019) Prem (2023)

Etiske principper og interessenter i AI-udviklingens livscyklus

Modeludvikling og træning

Interessenter

PROBLEM-FORMULERING OG FORRETNINGS-FORSTÅELSE	DESIGN-FASE	DATA-INDSAMLING, DATAFORSTÅELSE OG DATA FORBEREDELSE	MODELUDVIKLING OG TRÆNING	EVALUERING OG AFPRØVNING AF MODELLER	UDRULNING	DRIFT OG OVERVÅGNING

- **AI/ML-forskere** omdanner problemformuleringen til en prototype-AI-model.
- **AI-ingeniører** har ekspertise til at vælge de bedst egnede algoritmiske værktøjer til bestemte formål.
- **Dataforskere** er centrale i udviklingen af maskinlæringsmodeller og optimeringen af dem.
- **Etiske eksperter** er afgørende for at integrere etiske principper i udviklingsprocessen for at undgå utilsigtede konsekvenser.
- **Brugere i virksomheden** sikrer, at modellerne lever op til virksomhedens krav og forventninger.

(D. De Silva & D. Alahakoon, 2022), (J. Rochel & F. Évequoz, 2020)

Etiske principper og interessenter i AI-udviklingens livscyklus

Evaluering og afprøvning af modeller

Når modellen er udviklet og trænet, skal dens performance testes og evalueres:

- Evaluer modellen i forhold til foruddefinerede mål og succeskriterier.
- Gør klart rede for de underliggende antagelser og standarder, som datasættene er baseret på.
- Udfør tests med nye data data eller trin i processen for at sikre robusthed.
- Hvis resultaterne er utilfredsstillende, kan man overveje at ændre modelparametre, arkitektur eller endda datasættene bag datamodellen.
- Eventuelle mangler bør diskuteres åbent og ærligt med projektlederen og andre interessenter, da de kan have stor betydning for enkeltpersoner.
- Bliv enige om, hvorvidt I skal fortsætte med flere iterationer eller gå videre til implementeringsfasen.

Denne fase er afgørende for at sikre modellens pålidelighed og effektivitet, før den sættes i produktion.

Rochel (2023)/ Data Science Process Alliance

Etiske principper og interessenter i AI-udviklingens livscyklus

Evaluering og afprøvning af modeller

Etiske principper, der skal overvejes

PROBLEM-FORMULERING OG FORRETNINGS-FORSTÅELSE	DESIGN-FASE	DATA-INDSAMLING, DATAFORSTÅELSE OG DATA FORBEREDELSE	MODELUDVIKLING OG TRÆNING	**EVALUERING OG AFPRØVNING AF MODELLER**	UDRULNING	DRIFT OG OVERVÅGNING

Retfærdighed / gavnlighed / ikke-skadelighed

- **Etiske revisioner og målinger:**
  - Sikr overholdelse af etiske principper som retfærdighed, gavnlighed og ikke-skadelighed gennem revisioner og målinger.
- **Sikkerhed og modstandsdygtighed:**
  - Test modstandsdygtigheden over for angreb, og lav en fallback-plan for at sikre systemets sikkerhed.
- **Håndtering af negative påvirkninger og mulige risici:**
  - Implementer strategier for at minimere og kunne indrapportere problemer.
  - Etabler en mekanisme til at afhjælpe potentielle skadelige virkninger, hvis noget går galt.
  - Formuler risikovurderingsfaktorer med fokus på privatlivets fred, sociale konsekvenser og bias.
- **Forklarlighed:**
  - Forklar tydeligt de tekniske processer og menneskelige beslutninger, der er involveret i modeludviklingen.

Vær opmærksom på, at IT-medarbejdere, der er under pres for at færdiggøre løsninger til udrulning, kan have en øget risiko for at overse problemer.

Rochel (2023) / Prem (2019) / De Silva & Alahakoon (2022)

Etiske principper og interessenter i AI-udviklingens livscyklus

Evaluering og afprøvning af modeller

Interessenter

PROBLEM-FORMULERING OG FORRETNINGS-FORSTÅELSE	DESIGN-FASE	DATA-INDSAMLING, DATAFORSTÅELSE OG DATA FORBEREDELSE	MODELUDVIKLING OG TRÆNING	EVALUERING OG AFPRØVNING AF MODELLER	UDRULNING	DRIFT OG OVERVÅGNING

- **AI-ingeniører** fortolker resultaterne efter modelleringen.
- **Kvalitetssikringsteams (QA)** sikrer, at modellerne lever op til kvalitetsstandarder og er fri for fejl eller utilsigtede konsekvenser.
- **Etiske eksperter** sikrer etisk brug af AI-modeller.
- **Brugernes advokater/repræsentanter** sikrer, at AI-modellerne stemmer overens med brugernes forventninger og behov.
- **Legal and Compliance-teams** mindsker juridiske risici ved at sikre overholdelse af relevante regler.

(J. Rochel & F. Évequoz, 2020) / (Moore, 2023)

Etiske principper og interessenter i AI-udviklingens livscyklus

Udrulning

Når AI-løsningen er tilfredsstillende færdigudviklet, implementeres den i et produktionsmiljø for at løse problemer i den virkelige verden. Kompleksiteten i denne fase afhænger af selve løsningen.

- I første omgang introduceres løsningen for en mindre gruppe eksperter og brugere.
- Du kan enten integrere modellen med eksisterende systemer, skabe en applikation eller service, der bruger modellen, eller udnytte den i en offline-sammenhæng, f.eks. i en rapport til ledelsen.
- Etabler en feedback-loop til overvågning af performance og resultater, med efterfølgende iterationer, hvis det er nødvendigt.
- Implementer risikostyringsstrategier for at løse potentielle problemer under udrulningen.
- Planlæg løbende overvågning og vedligeholdelse i driftsfasen.
- Gennemgå hele projektet og udarbejd en endelig rapport, der dokumenterer processen og resultaterne.

Denne strukturerede tilgang sikrer, at AI-løsningen er effektivt integreret og i stand til at fungere efter hensigten i virkelige scenarier.

De Silva & Alahakoon (2022) / Data Science Process Alliance

Etiske principper og interessenter i AI-udviklingens livscyklus

Udrulning

Etiske principper, der skal overvejes

PROBLEM-FORMULERING OG FORRETNINGS-FORSTÅELSE | DESIGN-FASE | DATA-INDSAMLING, DATAFORSTÅELSE OG DATA FORBEREDELSE | MODELUDVIKLING OG TRÆNING | EVALUERING OG AFPRØVNING AF MODELLER | **UDRULNING** | DRIFT OG OVERVÅGNING

Autonomi

- **Menneskelig handlekraft**
  - Sørg for, at brugerne er klar over, hvornår beslutninger, indhold, råd eller resultater genereres af en AI. Fremhæv interaktionen med AI-løsningen for at forhindre overdreven afhængighed.
- **Menneskeligt tilsyn**
  - Sørg for, at der er menneskeligt tilsyn, som kan gribe ind, når det er nødvendigt, og sørg for, at beslutninger kan revurderes eller ændres af mennesker.

Retfærdighed

- Implementer stærke revisionsforanstaltninger for regelmæssigt at vurdere systemets ydeevne og overholdelse af etiske standarder.
- Etabler mekanismer, så brugerne kan sætte spørgsmålstegn ved beslutninger eller indgive klager, og sørg for beskyttelse af whistleblowere.

Unesco (2023) / Morley (2019)

Etiske principper og interessenter i AI-udviklingens livscyklus

Udrulning

Etiske principper, der skal overvejes

PROBLEM-FORMULERING OG FORRETNINGS-FORSTÅELSE | DESIGN-FASE | DATA-INDSAMLING, DATAFORSTÅELSE OG DATA FORBEREDELSE | MODELUDVIKLING OG TRÆNING | EVALUERING OG AFPRØVNING AF MODELLER | **UDRULNING** | DRIFT OG OVERVÅGNING

Forklarlighed

- Fremhæv vigtigheden af at forklare både de tekniske processer og de relaterede menneskelige beslutninger for brugerne.
- Tag fat på udfordringen med at forenkle kompleks menneskelig adfærd til værktøjer, der er letforståelige, og undgå teknisk jargon -> det øger tilliden og systemaccepten.
- Opretholde gennemsigtighed om AI-funktionaliteter og -begrænsninger.

Evaluering af langsigtede virkninger

- Vurder de langsigtede virkninger på samfundet og miljøet i denne fase.
- Overvej nødvendige forbedringer for at sikre, at AI-systemet bidrager positivt over tid.

Morley / UNESCO / Prem

Etiske principper og interessenter i AI-udviklingens livscyklus

Udrulning

Interessenter

- **AI/ML-ingeniører** omdanner prototype-AI-modellen til en implementeret service eller løsning, der er tilgængelig for alle interessenter og slutbrugere.
- **Sikkerhedseksperter** beskytter det implementerede system mod cybersikkerhedsrisici.
- **Slutbrugere** kan identificere eventuelle problemer eller begrænsninger, som måske ikke har været synlige under test i kontrollerede miljøer.
- **Brugerrepræsentanter** indsamler brugerfeedback om den implementerede AI-løsnings anvendelighed og performance og sikrer, at den implementerede løsning lever op til brugernes forventninger og behov.

(D. De Silva & D. Alahakoon, 2022) / (Junklewitz et al., 2023)

Etiske principper og interessenter i AI-udviklingens livscyklus

Drift og overvågning

Efter implementeringen kræver AI-løsningen løbende vedligeholdelse, opdateringer og løbende evaluering.

- Sørg for, at modellen fortsætter med at fungere som forventet ved regelmæssigt at overvåge dens effektivitet.
- Overvåg løbende slutbrugernes aktiviteter for at få indsigt i driften og opdage eventuelle problemer.
- Opdater jævnligt modellen med de nyeste data for at bevare dens nøjagtighed og relevans.
- Finpuds modellen baseret på brugerfeedback for at forbedre funktionalitet og ydeevne.
- Iterative opdateringer er ofte nødvendige og skal ses som en normal del af AI-livscyklussen, ikke som en fiasko.
- Overvåg investeringsafkast og andre relevante målinger baseret på de oprindeligt fastsatte mål.

Denne fase er afgørende for at sikre, at AI-løsningen forbliver effektiv og fortsat opfylder brugernes behov.

De Silva & Alahakoon (2022) / Data Science Process Alliance

Etiske principper og interessenter i AI-udviklingens livscyklus

Drift og overvågning

Etiske principper, der skal overvejes

PROBLEM-FORMULERING OG FORRETNINGS-FORSTÅELSE | DESIGN-FASE | DATA-INDSAMLING, DATAFORSTÅELSE OG DATA FORBEREDELSE | MODELUDVIKLING OG TRÆNING | EVALUERING OG AFPRØVNING AF MODELLER | UDRULNING | DRIFT OG OVERVÅGNING

Monitorering af etiske principper

- Gennemgå og vurder regelmæssigt AI-systemets overholdelse af etiske principper, idet der tages højde for mulige ændringer i datasæt og modelstrukturer.

Fortsat inddragelse af interessenter

- Opretholdet et aktivt engagement med brugere, organisationer og det bredere samfund for at indsamle løbende feedback.
- Etabler strukturerede mekanismer for at sikre regelmæssig og systematisk indsamling af viden og feedback.

Unesco (2023)/ HLEG (2019)

Etiske principper og interessenter i AI-udviklingens livscyklus

Drift og overvågning

Interessenter

- **DevOps- og IT-driftsteams** overvåger systemets performance og sikrer kontinuerlig tilgængelighed.
- **Instanser, som certificerer sikkerheden,** vurderer sikkerheden i den udviklede AI-løsning.
- **Ekspertpaneler, styregrupper eller tilsynsorganer** gennemgår projektet i forhold til tekniske og etiske aspekter.

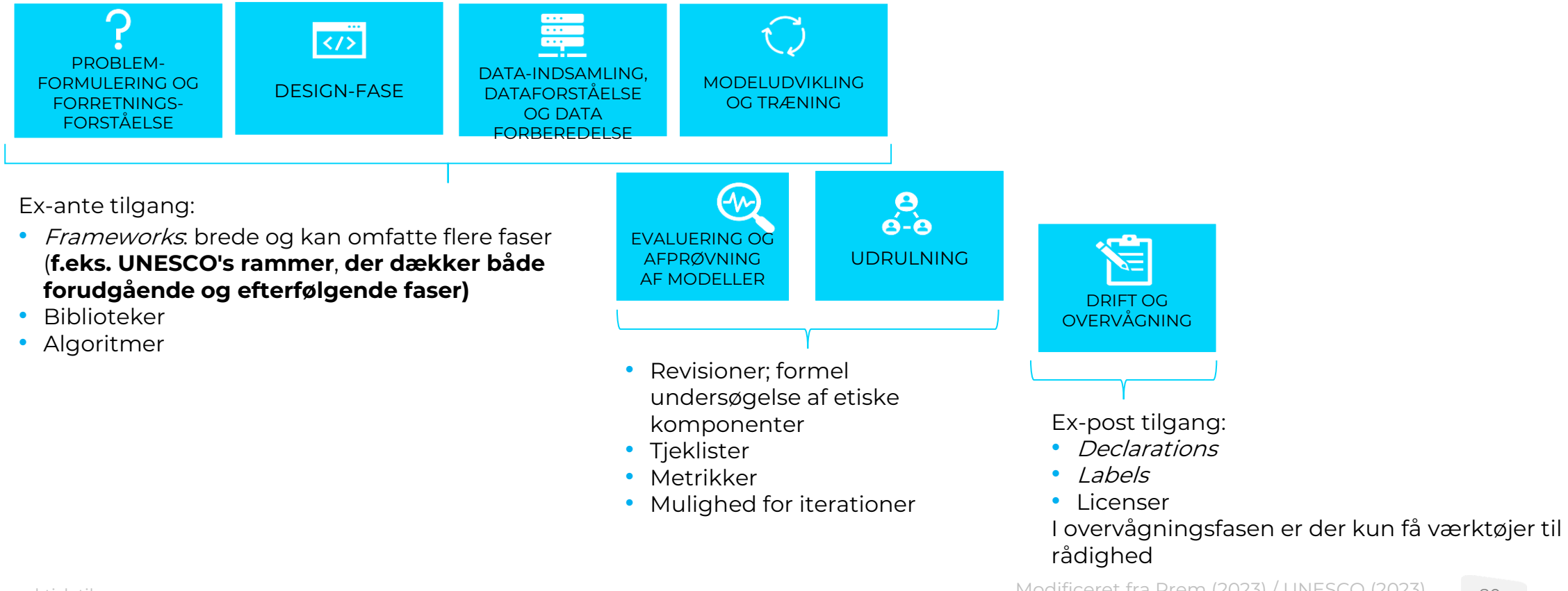
(D. De Silva & D. Alahakoon, 2022) / (Immersant Data Solutions, 2023) / (Miller, 2022)

Etiske principper og interessenter i AI-udviklingens livscyklus

Livscyklus for AI-udvikling

Værktøjer

De tilgængelige værktøjer til etiske overvejelser varierer meget afhængigt af AI-udviklingstrinnet. Mange tilgange kan være velegnede i alle trin, men nogle kun bruges i visse udviklingstrin. Nedenstående kategorisering er et groft billede af, hvordan de tilgængelige værktøjer kan opdeles mellem designstadierne.



Modificeret fra Prem (2023) / UNESCO (2023)

TAK

Medfinansieret af Den Europæiske Union

Finansieret af Den Europæiske Union. Synspunkter og holdninger, der kommer til udtryk, er udelukkende forfatterens/forfatternes og er ikke nødvendigvis udtryk for Den Europæiske Unions eller Det Europæiske Forvaltningsorgan for Uddannelse og Kulturs (EACEA) officielle holdning. Hverken den Europæiske Union eller EACEA kan holdes ansvarlig herfor.

2022-1-ES01-KA220-HED-000085257