EU’s etiske retningslinjer for pålidelig kunstig intelligens

ITB.DK

IT-BRANCHEN

EU’s etiske retningslinjer for pålidelig kunstig intelligens

EU har oprettet en række retningslinjer for at fremme brugen af pålidelig kunstig intelligens.

Brugen af AI skal ifølge EU’s retningslinjer grundlæggende leve op til disse tre ting gennem hele systemets levetid:

- Den skal være lovlig og overholde alle love og bestemmelser.
- Den skal være etisk og overholde etiske principper og værdier.
- Den skal være robust fra både et teknisk og et socialt perspektiv, da AI-systemer kan forårsage uforsættelig skade, selv med gode hensigter.

Evalueringsliste for brug af kunstig intelligens

Når man udvikler, udbreder og bruger kunstig intelligens, er det vigtigt, at man overvejer en lang række forskellige ting, så man sikrer, at det også sker forsvarligt og inden for lovens rammer.

Derfor har EU udarbejdet denne vejledning, som AI-aktører kan bruge for at sikre, at de overholder de etiske krav, som man forventer i Europa.

Vejledningen er lavet som en tjekliste og opdelt i syv dimensioner, som man altid bør undersøge, når man arbejder med AI:

1. Menneskelig udførelse og kontrol
2. Teknisk robusthed og sikkerhed
3. Overholdelse af privatlivets fred og datastyring
4. Gennemsigtighed
5. Diversitet, ikkediskrimination og retfærdighed
6. Samfunds- og miljømæssig velfærd
7. Ansvarlighed

Tanken med denne guideline er, at man bruger det som et samtaledokument eller tjekliste, der sikrer, at man overvejer alle relevante dimensioner, når man udvikler eller benytter AI-systemer.

Dette dokument er en forkortet og omskrevet udgave af EU’s oprindelige retningslinjer, som du kan læse hele den oprindelige udgivelse af på EU’s hjemmeside.

1 Menneskelig udførelse og kontrol

AI-systemer bør støtte menneskers autonomi og beslutningstagning. Det betyder, at AI-systemer bør fungere som støtte for et demokratisk, livfuldt og retfærdigt samfund ved at støtte brugernes handlinger og fremme grundlæggende rettigheder samt tillade menneskelig kontrol.

1.1 Grundlæggende rettigheder

1.1.1 Har du udført en konsekvensanalyse vedrørende borgernes grundlæggende rettigheder, hvor du har identificeret og dokumenteret potentielle afvejninger mellem de forskellige principper og rettigheder?

1.1.2 Er der risiko for, at AI-løsningen påvirker menneskers autonomi ved at gribe ind i slutbrugerens beslutningsproces på en utilsigtet måde?

1.1.3 Har du overvejet, om AI-systemet bør meddele brugerne, at en beslutning, indhold, rådgivning eller en konklusion er resultatet af en algoritmisk afgørelse?

1.1.4 Hvis AI-systemet omfatter en "chatbot" eller et andet samtalesystem, gør du så slutbrugerne opmærksomme på, at de interagerer med en ikke-menneskelig bot?

1.2 Menneskelig udførelse

1.2.1 Hvis AI-systemet er implementeret i en arbejdsproces, du så overvejet, om det er med til at forbedre eller styrke menneskelige færdigheder?

1.2.2 Har du truffet sikkerhedsforanstaltninger, så du forhindrer en overdreven tillid til eller afhængighed af AI-systemet i den pågældende arbejdsproces?

1.3 Menneskelig kontrol

1.3.1. Kan du beskrive, hvem der har den menneskelige kontrol, og hvornår og med hvilke værktøjer der kan ske menneskelig indgriben?

1.3.2 Har du indført specifikke mekanismer og foranstaltninger, der sikrer potentiel menneskelig kontrol, og som sikrer, at det er mennesker, der er ansvarlig for beslutningerne?

1.3.3 Har du truffet foranstaltninger, der gør det muligt at revidere og afhjælpe evt. problemer vedrørende styring af AI-autonomi?

1.3.4 Hvilken type detektions- og responsmekanismer har du etableret for at vurdere, om noget kan gå galt?

1.3.5 Har du etableret en "stopknap" eller procedure, der kan afbryde en operation på en sikker måde, hvis det bliver nødvendigt? Og afbryder den processen helt eller delvist, eller overlader den kontrollen til et menneske?

2 Teknisk robusthed og sikkerhed

At et system er teknisk robust betyder, at det er bygget, så det ikke bare bryder sammen, og at man minimerer skaden, hvis uheldet alligevel sker. Dette inkluderer også overvejelser om det tekniske setup i driftsmiljøet samt at sikre, at folk (eller AI'er) ikke kan modarbejde systemet.

2.1. Modstandsdygtighed over for angreb og sikkerhed

2.1.1 Har du overvejet de forskellige sårbarheder (herunder dataforurening, infrastruktur og cyberangreb) og de potentielle angreb, som AI-systemet kan blive udsat for?

2.1.2 Har du indført foranstaltninger, der kan sikre AI-systemets modstandsdygtighed overfor angreb?

2.1.3 Har du vurderet, hvordan dit system skal opføre sig i uventede situationer?

2.1.4 Har du overvejet, om løsningen kan have dobbeltanvendelse (at det kan bruge til både fredelige og militære formål), og har du i så fald lavet forebyggende foranstaltninger mod dette?

2.2 Fallback-plan og generel sikkerhed

2.2.1 Har du sørget for en tilstrækkelig fallback-plan, hvis systemet udsættes for fjendtlige angreb, eller andre uventede situationer (f.eks. tekniske omskiftningsprocedurer eller krav om en menneskelig operatør, inden proceduren fortsættes)?

2.2.2 Har du indført en proces til måling og vurdere af risici og sikkerhed samt overvejet det risikoniveau AI-systemet giver anledning til ved en specifik anvendelse?

2.2.3 Har du identificeret potentielle sikkerhedsrisici, hvis løsningen bruges på anden måde end tiltænkt – f.eks. ved uforsætlig eller ondsindet misbrug? Og er der fastlagt en plan for minimering af disse risici?

2.2.4 Har du overvejet en forsikringsstrategi, der kan håndtere potentielle skader fra AI-systemet?

2.2.5 Hvis der er risiko for, at AI-systemet kan forårsage skade for brugere eller tredjemand, har du så vurderet sandsynligheden for den potentielle skade, den berørte målgruppe og alvoren?

2.2.6 Har du estimeret, hvad følgevirkningerne er på forskellige fejl, og har du i den forbindelse defineret og testet tærskler og processer, der udløser alternative planer eller fallback-planer?

2.3 Nøjagtighed

2.3.1 Har du vurderet, hvilket niveau og hvilken definition af nøjagtighed der kræves i forbindelse med AI-systemet samt hvordan nøjagtigheden måles og kontrolleres?

2.3.2. Har du indført foranstaltninger der sikrer, at de anvendte data er fuldstændige og ajourførte?

2.3.3 Har du vurderet den skade, der kan forårsages, hvis systemet foretager unøjagtige forudsigelser?

2.3.4 Hvis der foretages unøjagtige forudsigelser, har du indført processer for at løse problemet?

2.4 Pålidelighed og reproducerbarhed

2.4.1 Overvåger, dokumenterer og tester du, om AI-systemet opfylder de fastsatte mål, formål og tiltænkte anvendelser?

2.4.2 Har du testet reproducerbarheden – altså at resultaterne kan genskabes i systemet eller eksternt?

2.4.3 Har du indført processer, der beskriver, hvornår AI-systemet svigter under bestemte omstændigheder?

2.4.4 Har du indført en mekanisme eller kommunikation, der forsikrer slutbrugerne om AI-systemets pålidelighed?

3 Overholdelse af privatlivets fred og datastyring

Privatlivets fred er en grundlæggende ret, der kan udfordres af AI-systemer. Beskyttelser af privatlivets fred kræver datastyring, så man altid sikrer relevansen, kvaliteten og integriteten af de anvendte data.

3.1 Respekt for privatlivets fred og databeskyttelse

3.1.1 Har du etableret en mekanisme, der giver andre mulighed for at give dig besked, hvis de oplever problemer vedrørende privatlivets fred eller databeskyttelse (både under oplæring og drift)?

3.1.2 Har du vurderet typen og omfanget af data i dit datasæt (f.eks. om de indeholder personoplysninger)?

3.1.3 Har du overvejet, hvordan du kan udvikle og oplære AI'en uden eller med minimal anvendelse af potentielt følsomme eller personlige data?

3.1.4 Har du mekanismer der giver besked og kontrol over personoplysninger (f.eks. gyldigt samtykke og mulighed for tilbagekaldelse)?

3.1.5 Har du overvejet at benytte f.eks. kryptering, anonymisering og aggregering for at beskytte privatlivets fred?

3.1.6 Hvis din virksomhed har en databeskyttelsesansvarlig, er denne så blevet involveret i en tidlig fase af processen?

3.2 Datakvalitet og -integritet

3.2.1 Har du etableret kontrolmekanismer for dataindsamling, -lagring, -behandling og -anvendelse?

3.2.2 Har du tilpasset den daglige datahåndtering og -styring til relevante protokoller eller standarder (f.eks. ISO og IEEE)?

3.2.3 Har du vurderet, hvor meget kontrol du har med kvaliteten af eksterne datakilder?

3.2.4 Har du indført processer for at sikre kvaliteten og integriteten af dine data?

3.2.5 Har du overvejet, hvordan du sikrer, at dine datasæt ikke er blevet kompromitteret eller hacket?

3.3 Adgang til data

3.3.1 Har du vurderet, hvem der kan få adgang til brugerdata og under hvilke omstændigheder?

3.3.2 Har du sørget for, at disse personer besidder kompetencerne, er kvalificerede og har behov for at få adgang til dataene?

3.2.3 Har du sørget for en overvågningsfunktion der registrerer, hvornår, hvor, af hvem og til hvilke formål data er tilgået?

4 Gennemsigtighed

Dette krav hænger tæt sammen med princippet om forklarlighed og omfatter gennemsigtighed af f.eks. data, systemet og forretningsmodellerne.

4.1 Sporbarhed

4.1.1 Har du indført foranstaltninger, der sikrer sporbarhed?

4.1.2 Hvis der er tale om design og udvikling af en regelbaseret AI-løsning, kan du så dokumentere programmeringsmetoden eller den måde, hvorpå modellen er bygget på? Hvis der er tale om design og udvikling af en læringsbaseret AI-løsning, kan du så dokumentere, hvilke inputdata der er indsamlet, og hvordan de er udvalgt?

4.1.3 Hvis der er tale om test og validering af en regelbaseret AI-løsning, kan du så dokumentere de scenarier eller cases, der er anvendt? Hvis der er tale om test og validering af en læringsbaseret AI-løsning, kan du så dokumentere de data, der er anvendt?

4.1.4 Kan du dokumentere resultaterne af eller afgørelser truffet af algoritmen, også i forhold forskellige cases (f.eks. for andre undergrupper af brugere)?

4.2 Forklarlighed

4.2.1 Har du vurderet, hvorvidt AI-systemets afgørelser og resultater kan forstås?

4.2.2 Kan du forklare, hvorfor et system kom frem til et bestemt resultat, så det er forståeligt for alle brugere?

4.2.3 Kan du forklare, hvorfor netop dette system blev implementeret på dette specifikke område?

4.2.4 Har du vurderet, i hvilken udstrækning systemets afgørelse påvirker organisationens beslutningsprocesser?

4.2.5 Har du vurderet forretningsmodellen for systemet (f.eks. hvordan skaber det merværdi for organisationen)?

4.2.6 Har du forsøgt at anvende den enkleste og mest fortolkningsvenlige model?

4.2.7 Har du vurderet, om du kan analysere dine oplærings- og testdata samt ændre og ajourføre dem over tid?

4.3 Kommunikation

4.3.1 Har du meddelt slutbrugerne, at de interagerer med et AI-system og ikke med et andet menneske?

4.3.2 Informerer du klart og tydeligt til brugerne om årsagerne til og kriterierne for AI-systemets resultater samt kommunikeret om potentielle eller opfattede risici, f.eks. skævheder?

4.3.3 Har du etableret processer, der tager hensyn til brugernes feedback?

4.3.4 Har du præciseret, hvad formålet med AI-systemet er, og hvem eller hvad der kan få gavn af produktet/tjenesten?

4.3.5 Har du klart og tydeligt kommunikeret AI-systemets karakteristika, begrænsninger og potentielle mangler?

4.3.6 Har du overvejet kommunikation og gennemsigtighed. over for andre målgrupper, tredjemand eller offentligheden?

5 Diversitet, ikkediskrimination og retfærdighed

For at opnå pålidelig kunstig intelligens skal vi sikre inklusion og diversitet i hele livscyklussen for AI-systemet. Ud over at tage hensyn til og inddrage alle berørte interessenter i løbet af processen omfatter dette også sikring af lige adgang gennem inklusive designprocesser og ligebehandling.

5.1 Undgåelse af urimelig skævhed

5.1.1 Har du sikret procedurer, der forhindrer, at urimelig skævhed skabes eller forstærkes i AI-systemet, både med hensyn til brugen af inputdata og med hensyn til algoritmedesignet?

5.1.2 Har du taget hensyn til diversiteten og repræsentativiteten af brugere i dataene? Og har du testet for specifikke populationer eller problematiske anvendelser?

5.1.3 Har du indført processer for test og overvågning af potentielle skævheder – både under udvikling, udbredelse og anvendelse af systemet?

5.1.4 Har du etableret en mekanisme, der giver andre mulighed for at udpege problemer med skævhed, diskrimination eller dårlig ydeevne i forbindelse med AI-systemet?

5.1.5 Har du vurderet, om der kan ske forskel på afgørelser på trods af de samme grundlæggende betingelser, hvad dette i så fald kan skyldes og hvilken påvirkning det vil have?

5.1.6 Har du etableret mekanismer til at sikre retfærdighed i dine AI-systemer, og har du fastlagt en passende arbejdsdefinition af "retfærdighed", som du anvender, når du designer AI-systemer?

5.2 Tilgængelighed og universelt design

5.2.1 Har du sikret, at AI-systemet understøtter et bredt spektrum af individuelle præferencer og funktioner? F.eks. om løsningen kan anvendes af personer med særlige behov eller handicap.

5.2.2 Har du sikret, at oplysninger om AI-systemet også er tilgængelige for brugere af teknologiske hjælpemidler?

5.2.3 Er det team, der er involveret i opbygningen af løsningen, repræsentativt for din målgruppe af brugere?

5.2.4 Har du vurderet, om der kan være personer eller grupper, som kan blive uforholdsmæssigt berørt af negative konsekvenser?

5.3 Inddragelse af interessenter

5.3.1 Har du overvejet at involvere forskellige interessenter i udviklingen og anvendelsen af AI-systemet?

5.3.2 Har du informeret og inddraget de berørte medarbejdere og deres repræsentanter i forbindelse med udviklingen og implementeringen af AI -systemet i din organisation?

6 Samfunds- og miljømæssig velfærd

Ideelt bør kunstig intelligens anvendes til fordel for alle mennesker, herunder fremtidige generationer. Derfor bør AI-systemer også understøtte bæredygtighed og økologisk ansvarlighed ligesom forskning i AI, der omhandler områder af global relevans, bør fremmes.

6.1 Bæredygtig og miljøvenlig kunstig intelligens

6.1.1 Måler du miljøvirkningen af AI-systemet både under dens udvikling, udbredelse og anvendelse (f.eks. datacentres energiforbrug, energitype, der anvendes af datacentre osv.)?

6.1.2 Har du etableret foranstaltninger til at nedbringe miljøvirkningen af AI-systemets livscyklus?

6.2 Social indvirkning

6.2.1 Har du vurderet, om AI-løsningen får mennesker til at knytte sig til eller have empati over for systemet? Og har du i den forbindelse sikret, at AI-systemet klart signalerer, at dets sociale interaktion er simuleret, og at det ikke har kapacitet til at "forstå" og "føle"?

6.2.2 Har du sikret, at de sociale indvirkninger af AI-systemet opfattes klart (f.eks. om der er risiko for tab af arbejdspladser eller nedkvalificering af arbejdsstyrken)?

6.3 Samfund og demokrati

6.3.1 Har du vurderet, hvordan AI-løsningen påvirker det bredere samfund f.eks. de potentielt indirekte berørte aktionærer?

7. Ansvarlighed

Kravet om ansvarlighed supplerer ovennævnte krav og hænger tæt sammen med princippet om retfærdighed. Det kræver, at der indføres mekanismer, som sikrer ansvaret og ansvarligheden for AI-systemer og deres resultater, både før og efter implementeringen.

7.1 Mulighed for revision

7.1.1 Har du indført mekanismer, der gør det lettere at revidere systemet for interne og/eller uafhængige aktører, f.eks. ved at sikre sporbarhed og registrering af processer og resultater?

7.2 Minimering og rapportering af negative virkninger

7.2.1 Har du udført en risiko- eller konsekvensanalyse af AI-systemet, som tager hensyn til de forskellige interessenter, der påvirkes direkte og indirekte?

7.2.2 Har du indført oplærings- og uddannelsesrammer vedr. ansvarlighed overfor det involverede team og andre medarbejdere. Og omfatter oplæringen også undervisning i de relevante lovgivningsmæssige rammer?

7.2.3 Er der etableret procedurer, som tredjeparter (f.eks. leverandører, forbrugere og distributører/forhandlere) kan bruge til at rapportere potentielle sårbarheder, risici eller skævheder i AI-systemet/anvendelsen?

7.3 Dokumentation af afvejninger

7.3.1 Har du dokumenteret og etableret en mekanisme til at identificere relevante interesser og værdier, der påvirkes af AI-systemet?

7.3.2 Har du beskrevet, hvilken proces du bruger til at træffe afgørelse om afvejninger mellem forskellige interesser og værdier? Har du sikret, at afgørelsen om afvejning blev dokumenteret?

7.4 Klageadgang

7.4.1 Har du etableret processer, der giver adgang til klage i tilfælde af skade eller negative indvirkninger?

7.4.2 Har du indført tydelige mekanismer der giver slutbrugere og tredjeparter oplysninger om klageadgang?