FINANSTILSYNET

NOTAT

Finanstilsynet

29. maj 2024

God praksis ved brug af kunstig intelligens

De væsentlige gennembrud indenfor generativ kunstig intelligens (AI) i efteråret 2022, særligt den mulige brug af ChatGPT og lignende tjenester, har rykket AI op på dagsordenen i samfundet. Det gælder også internt i finansielle virksomheder, der kan udnytte teknologien til at vinde markedsandele eller på anden måde optimere. Disse tjenester gør det muligt for alle, også brugere uden særlige tekniske forudsætninger, at bruge AI i deres hverdag. Ikke mindst, eller måske særligt, derfor kan det være nødvendigt at gøre noget aktivt for at undgå, at en frygt for at blive overhalet kan komme til at overskygge fokus på god governance og risikostyring samt på den sunde virksomhedskultur og de etiske overvejelser, der er en forudsætning for betryggende anvendelse.

I det lys vurderer Finanstilsynet, at anbefalingerne fra *God praksis ved brug af superviseret machine learning* (maskinlæring) i den finansielle sektor, der blev offentliggjort i 2019[1], med fordel kan konkretiseres yderligere. Anbefalingerne i dette god-praksis-papir skal derfor læses i forlængelse af 2019-papiret.

Papiret fra 2019 indeholder Finanstilsynets anbefalinger på i alt ni konkrete områder, som virksomheder med fordel kan overveje, i takt med at brugen af maskinlæring, eller bredere – AI, stiger. I takt med, at AI rykker fra tegnebrættet til produktionsmiljøet, ser Finanstilsynet et behov for endnu engang at pointere, at tilgangen til brug af teknologien, uagtet dens store potentiale, bør have fokus på risici.

God praksis-papiret fra 2019 tog udgangspunkt i et konkret testforløb i den regulatoriske sandkasse, FT Lab, af superviseret maskinlæring via et dybt neuralt netværk. Af den grund valgte Finanstilsynet at udgive det oprindelige god praksis-papir for brug af superviseret maskinlæring, omend de emner og forslag til god praksis, som 2019-papiret indeholder, efter Finanstilsynets mening beskriver god praksis for brug af AI generelt.

[1] https://www.finanstilsynet.dk/Tilsyn/Information-om-udvalgte-tilsynsomraader/Fintech/Machine_learning_10719

Finanstilsynets god praksis-papir fra 2019 oplistede en række forslag til god praksis for brug af maskinlæring med fokus på følgende områder:

- Formål med brug af superviseret machine learning og beskrivelse af modellen
- Governance (modeludvikling, -anvendelse og -opdatering), politikker og forretningsgange
- Datahåndtering
- Træning af modellen
- Performance og robusthed
- Ansvarlighed (accountability)
- Forklarlighed (explainanbility)
- Dataetik, skævhed i data (bias) og rimelighed (fairness)
- Gennemsigtighed (transparancy).

Dette papir tager afsæt i en række møder med AI-eksperter. Møderne havde fokus på AI i bredere og mere generel forstand end alene maskinlæring. Det er derfor også det fokus, der er anlagt i papiret.

Afsnittet om *governance* er i dette papir opdelt i to underafsnit med fokus på henholdsvis *organisationen* og *modellen.* Også afsnittet om *forklarlighed* bliver genbesøgt.

Papiret er udarbejdet sideløbende med, at EU's kunstig intelligens-forordning (AI-forordningen) blev udformet. Formålene med AI-forordningen og dette papir er dog forskellige. AI-forordningen dækker teknologianvendelsen bredt i samfundet og introducerer konkret regulering rettet mod selskaber, der bruger eller sælger produkter baseret på AI. Forordningen fokuserer på at beskytte EU-borgeres fundamentale rettigheder i mødet med AI, mens dette papir fokuserer på risici for finansielle virksomheder, der bruger AI.

God praksis-papiret har ikke reguleringsmæssig karakter og vil ikke i sig selv danne grundlag for tilsynsmæssige reaktioner. Som led i Finanstilsynets 2025-strategi, der bl.a. skal understøtte *betryggende brug af teknologi samt nye forretningsmodeller,* er formålet med papiret at gøre de finansielle virksomheder opmærksomme på områder, hvor brug af AI kan medføre et øget behov for risikomitigerende handlinger.

Papiret forudsætter, at finansielle virksomheder grundlæggende har en betryggende IT-udvikling. Papiret har derfor fokus på at øge virksomhedernes bevidsthed om, at nye værktøjer kan skabe nye risici, og at disse nye risici forventeligt vil kræve en ny og anden form for håndtering. Derfor må den enkelte virksomhed overveje, om en opdatering af etablerede processer for risikostyring, herunder i forbindelse med IT-udvikling, er påkrævet.

Virksomhederne bør gøre sig en række etiske overvejelser ved brug af AI. De er ikke omfattet af dette papir, men Finanstilsynets overvejelser om etik findes i Finanstilsynets rapport om dataetik ved brug af AI i den finansielle sektor, der blev udgivet den 13. november 2023[2].

Dette papir er rettet mod alle regulerede udbydere af finansielle tjenester. Flere af disse virksomheder er omfattet af Bekendtgørelse om ledelse og styring af pengeinstitutter m.fl. eller Bekendtgørelse om ledelse og styring af forsikringsselskaber m.v., hvilket konkret betyder, at nogle af papirets forslag til god praksis for disse virksomheder vil være reelle krav.

Finanstilsynet har holdt møder med en række repræsentanter fra kreditinstitutter, investeringsselskaber, markedspladser, forsikrings- og pensionsselskaber samt betalingsinstitutter og rådført sig med eksperter og akademikere. Alle med indgående kendskab til brugen af AI i og udenfor den finansielle sektor. Fokus på møderne var brugen af AI, hvilke risici der kan opstå, og hvilke foranstaltninger virksomhederne har implementeret for at imødegå disse risici. Alle, der har bidraget til arbejdet, har haft mulighed for at give bemærkninger inden offentliggørelsen af papiret.

[2] https://www.finanstilsynet.dk/Tal-og-Fakta/Rapporter/2023/AI-i-den-finansielle-sektor_161123

1. Governance

En virksomhed, der bruger eller planlægger at bruge AI til at udføre sine aktiviteter, bør sikre sig, at organisationen har de bedst mulige betingelser for at implementere teknologien på en betryggende måde. Indretningen af organisationen bør sikre, at fordele bliver udnyttet, samtidig med at relevante risici, herunder også IT-sikkerhedsmæssige aspekter, bliver håndteret. Den enkelte virksomheds organisatoriske setup afhænger af virksomhedens forretningsområde og af den konkrete brug af teknologien. Der er med andre ord ikke én universel tilgang, der er hensigtsmæssig for alle.

Brugen af AI varierer på tværs af virksomheder, f.eks. i forhold til, om modellerne bruges direkte i forretningen eller til understøttende aktiviteter, såsom kvalitetskontrol eller simpel sortering af mails.

Nogle finansielle virksomheder bruger AI-modeller til at udfordre etablerede og myndighedsgodkendte modeller, som er baseret på mere klassiske statiske metoder. Når forskellen på performance mellem den godkendte model og den såkaldte udfordrermodel bliver tilpas stor, kan det indikere, at den etablerede model skal opdateres. I sådanne tilfælde påvirker AI ikke nødvendigvis direkte de etablerede processer som primært omhandler de statiske myndighedsgodkendte modeller, men teknologien understøtter en løbende test.

I andre sammenhænge bliver AI brugt direkte i etablerede forretningsgange, som f.eks. modeller til monitorering af transaktioner eller handler, og til automatisk videresendelse af mails fra en hovedpostkasse.

Krav og forventninger til virksomhedens risikostyring vil variere, alt afhængig af hvor stor effekt brugen af modellen har for virksomheden selv eller dens kunder, og om AI f.eks. bliver brugt direkte eller som en udfordrermodel. Jo større indflydelse AI har, jo mere forventer Finanstilsynet, at virksomheden har foranstaltninger til at identificere og håndtere risici.

System til virksomhedens overblik

De enkelte virksomheder har indrettet sig vidt forskelligt for effektivt at kunne identificere og håndtere risici fra brugen af AI. Flere har konkret fokus på AI i relevante politikker, mens andre har udarbejdet nye politikker for den konkrete teknologi. Nogle virksomheder har nedsat ledelsesmæssige organer, som tager stilling til udvikling og brug af AI, når effekten for deres kunder eller for virksomheden bliver tilpas stor. En del har også sikret sig, at de har overblik over, hvilke af deres modeller, indkøbte eller internt udviklede, der bruger AI. Et sådant overblik kan f.eks. etableres ved, at virksomhedens brug af modeller baseret på AI fremgår af et register over virksomhedens samlede brug af modeller, som alle virksomheder i øvrigt bør have. Den konkrete tilgang afhænger af, hvad der er mest hensigtsmæssigt for den enkelte virksomhed, men generelt kan et register over virksomhedens modeller bidrage til at sikre

virksomhedens overblik og indgå i organisationens generelle håndtering af den samlede modelrisiko.

Stillingtagen til særlige risici ved AI-baserede modeller

De virksomheder, der deltog i møderne med Finanstilsynet, har iværksat forskellige risikoreducerende tiltag. Nogle centraliserer kontrollen hos særligt kompetente medarbejdere, f.eks. domæneeksperter, som vurderer forskellige aspekter. Det kan f.eks. være hensigtsmæssigheden af en model i forhold til dens anvendelsesområde eller påvirkning og risici i tilknytning til en model set i forhold til internt fastsatte grænser. Andre virksomheder har interne politikker for risikostyring af modeller, men decentral vurdering af selve risikoen. Generelt bør en virksomhed tage stilling til, hvordan den sikrer håndtering af særlige risici tilknyttet modeller baseret på AI, og hvilke interne ressourcer den har til opgaven. Dette bør ske i forbindelse med, at den enkelte virksomhed fastlægger sin tilgang til risikohåndtering.

Etableret tilgang til risikoanalyse

Flere virksomheder bruger såkaldte tiers, der kategoriserer modeller ud fra en risikobetragtning, til at vurdere, hvor mange ressourcer de bør afsætte til risikoreducerende tiltag for den enkelte model. Tiering er en metode, hvor potentielle risici ved en model reduceres til en enkelt score. Denne score afgør, hvor modellen passer ind i virksomhedens governance. Sagt med andre ord indikerer placeringen i et konkret tier, hvor mange ressourcer virksomheden skal afsætte til at sikre, at brugen af modellen ikke overskrider virksomhedens risikotolerance. Tiering er blot ét eksempel på en metode til at kategorisere modeller. Hvilken metode, den enkelte virksomhed vælger, er ikke afgørende. Det væsentlige er, at virksomheden overhovedet har en procedure for at vurdere, hvor risikobetonet en model og dens anvendelsesområde er, så virksomheden kan fastsætte passende risikomitigerende tiltag. Proceduren for modeller baseret på AI kan være den samme, som virksomheden allerede bruger for sine øvrige modeller. Under alle omstændigheder er en etableret tilgang til risikoanalyse og klassificering af internt såvel som eksternt udviklede modeller grundlæggende for at kunne identificere og håndtere risici, herunder risici i tilknytning til AI. Det kan være relevant at inddrage emner som modellens kompleksitet og væsentlighed, samt hvor afhængig virksomheden er af en konkret models output i udførelsen af en funktion, som modellen understøtter eller udfører.

Identifikation og afgrænsning af risici

I takt med at brugen af AI stiger, forventer Finanstilsynet, at den enkelte virksomhed løbende forholder sig til, om dens tilgang til identifikation og afgrænsning af risici er tilstrækkeligt dækket af eksisterende politikker, om disse skal tilrettes, eller om det er nødvendigt at udarbejde en særskilt politik for brugen af AI. Det gælder f.eks. også, at virksomheden bør overveje, om brugen af AI

kan medføre så væsentlige risici, at virksomheden løbende bør rapportere til ledelsen og bestyrelsen.

Forankring af ansvaret hos konkrete medarbejdere eller enheder

Det vigtigste er, at virksomheden er i stand til at identificere og forholde sig til de risici, der er forbundet med brug af teknologien i den konkrete virksomhed. Virksomheden bør i den forbindelse have fokus på at forankre ansvaret for konkrete modeller, herunder for kontrol, hos konkrete medarbejdere eller enheder. Afhængig af væsentlighed, bør ansvaret desuden ligge hos andre end modeludvikleren. Uanset hvad virksomheden bruger, eller påtænker at bruge, AI til, er det nødvendigt at overveje tiltag, der kan sikre, at teknologien bliver implementeret og brugt på betryggende vis.

God praksis for brug af kunstig intelligens er, at virksomheden:

- har et system til at skabe overblik over sin brug af AI, evt. ved hjælp af det register, der bør være for modelanvendelse generelt
- tager stilling til, hvordan den sikrer håndtering af særlige risici, når den fastlægger sin tilgang for risikohåndtering, og hvilke interne ressourcer der er til opgaven
- har etableret en tilgang til risikoanalyse og klassificering af brugen af AI for både internt og eksternt udviklede modeller
- løbende forholder sig til, om dens tilgang til identifikation og afgrænsning af risici er tilstrækkelig dækket af eksisterende politikker
- sikrer, hvor det er relevant i forhold til modellens væsentlighed, at ansvaret for konkrete modeller, herunder kontrol, ligger hos konkrete medarbejdere eller enheder i organisationen, som ikke indgår eller har indgået i modeludviklingen.

2. Modelstyring

Med udgangspunkt i sin overordnede risikovurdering af brug af AI kan virksomheden forholde sig til, hvordan den håndterer eventuelle risici ved konkrete modeller. Virksomhedens eksisterende tiltag til håndtering af modelrisici kan være et godt udgangspunkt.

Bevidsthed om forskellen på AI-baserede og øvrige modeller

Flere virksomheder har gjort opmærksom på, at modeller, der bygger på AI, ikke nødvendigvis altid adskiller sig markant fra mere klassiske modeller. En virksomhed, der vil bruge AI, kan derfor med fordel identificere, hvor i en models livscyklus virksomhedens traditionelle metode til at håndtere modelrisici vil kunne blive udfordret af brugen af AI. Den enkelte virksomhed bør derfor forholde sig til, om og hvordan dens modeller baseret på AI adskiller sig fra andre modeller, og hvordan det påvirker virksomhedens styring af disse. En sådan risikoanalyse er forudsætningen for, at virksomheden kan håndtere eventuelle udfordringer. Det er ikke nødvendigvis givet på forhånd, hvor en ny model vil adskille sig. Brugen af AI vil muligvis give anledning til at opdatere forholdsregler i forbindelse med f.eks. udvikling, test, validering, ibrugtagning, gentræning, udfasning og lignende. Et eksempel, der går igen blandt virksomhederne, er, at modeller baseret på AI bl.a. adskiller sig fra mere klassiske modeller ved, at det ofte er relevant med hyppigere gentræning.

Frekvensen af gentræning kan fastsættes på flere forskellige måder, som hver især kan være hensigtsmæssige, afhængigt af den enkelte model og virksomhed. Flere virksomheder har f.eks. planlagt faste intervaller. Andre har fastlagt niveauer for, hvor langt en models estimater må være fra realiserede værdier, før modellen skal gentrænes. Endnu andre virksomheder har valgt at tage løbende stilling til, om den problemstilling, modellen er sat til at beskrive, har ændret sig radikalt. Det kan f.eks. være større hændelser som ændrede renteniveauer, geopolitiske spændinger, recession eller en pandemi. Det kan også være mindre hændelser som ændringer i datainput eller datakilder. Nogle virksomheder benytter flere sideløbende tilgange.

Plan for gentræning af modeller

Når processer bliver automatiserede med f.eks. brug af AI, kan det føre til ændret adfærd hos dem, der bruger systemet, eller dem, systemet bliver anvendt på, hvis de er bekendte med det. Det kan være ansatte, der administrerer systemet, eller de forbrugere eller kunder, som modellen er baseret på. En ændring i adfærd kan føre til, at modellens datagrundlag ikke længere afspejler den virkelighed, som systemet opererer i. En sådan risiko bør mitigeres, f.eks. gennem hyppigere gentræning. Grundlæggende bør virksomheden dokumentere, at den har taget stilling til, hvordan en konkret model forventes gentrænet for at holde den ajour. Det er især relevant, hvis modellen er udviklet af en tredjepart. Med afsæt i konkrete modellers risici, bør virksomheden have en plan for, hvor ofte og hvordan dens modeller skal gentrænes.

Det gælder for internt udviklede såvel som for eksternt indkøbte modeller. Endeligt bør virksomheden sikre regelmæssig validering.

Det er i sidste ende op til den enkelte virksomhed at vurdere, hvilke parametre den vil lægge til grund for frekvensen og metoden for gentræning. Den proces, der går forud for en konkret plan for gentræning af en model, og som f.eks. involverer indsamling af relevante data for beslutningen, giver virksomheden et godt indblik i, om formålet med modellen er retvisende beskrevet, og et godt overblik over modellens løbende udvikling.

Tilstrækkelige interne ressourcer

Gentræning af en kompleks model kan potentielt forandre modellen i en sådan grad, at det ikke længere kan betragtes som en mindre opdatering af en tidligere version. Der kan reelt være tale om en helt ny model. Derfor kan det give mening allerede på et tidligt stadie i modellens livscyklus at tænke over, hvordan den skal trænes og ikke mindst gentrænes, og hvilke ressourcer virksomheden bør have til rådighed på de aktuelle tidspunkter. Det kan f.eks. være, at domænekendskab i forbindelse med gentræning er vigtig for konkrete modelspecifikationer. Det vil i samme ombæring være en fordel også at indtænke modelvalidering. Hvis virksomheden kun har begrænsede ressourcer til modelvalidering, bør virksomheden muligvis prioritere AI-typer, der er mindre ressourcekrævende at validere. Virksomheden bør forholde sig til, om den har tilstrækkelige interne ressourcer, herunder til modelvalidering.

Validering af modeller baseret på AI bør som udgangspunkt følge samme tilgang, som virksomheden i øvrigt følger. F.eks. kan virksomheder med en separat valideringsenhed placere ansvaret der. Det betyder i praksis, at virksomheden bør sikre, at validering sker uafhængigt af udvikling, i det omfang virksomheden vurderer, at modellen er tilstrækkeligt væsentlig. Det bør altså som udgangspunkt ikke være den samme person eller gruppe, der både udvikler og efterfølgende validerer en given model hvis den vurderes tilstrækkeligt væsentlig.

For at sikre tilstrækkelig validering af brug af væsentlige modeller vil det som udgangspunkt være nødvendigt at have medarbejdere med de påkrævede tekniske og forretningsmæssige forudsætninger til at stå for en reelt kritisk kontrol. En kritisk kontrol af virksomhedens modeller er væsentlig for at mitigere risici, men i lyset af at flere har udfordringer med at rekruttere medarbejdere med tilstrækkelige kompetencer til selve udviklingen af modellerne, kan det potentielt også vise sig vanskeligt at rekruttere medarbejdere med lignende kompetencer til modelvalidering. Dette kan indikere, at virksomheden bør undgå særligt komplekse modeller. Virksomheden bør desuden allokere selvstændige ressourcer til den konkrete opgave med modelvalidering. Det vil ellers være oplagt at lade opgaven tilfalde modeludvikleren, som dermed vil skulle kontrollere eget arbejde, hvilket, særligt for modeller med større risici, ikke er hensigtsmæssigt.

Nogle virksomheder har fremhævet, at kompleksiteten kan stige yderligere, hvis resultater fra en model baseret på AI bliver input til andre modeller. Problematikken bliver yderligere forstærket, hvis komplekse og mindre forklarlige modeller indgår. Det kan f.eks. hurtigt blive svært at forudse og gennemskue, hvad en ellers simpel opdatering eller gentræning af en bagvedliggende model kan komme til at betyde for modeller længere nede i kæden. Derfor bør virksomhederne have metoder til at holde styr på sådanne sammenhænge, inden output fra én model bruges som input til andre.

Politik for styring og sikring af modelversioner

Hvis det bliver relevant at genskabe en tidligere version af en konkret kompleks model, kan det være nødvendigt at tilgå dokumentation – ikke bare for modellens parametre, men f.eks. også for, hvordan disse parametre blev valgt, hvilket datagrundlag der lå til grund, og for processen fra datainput til modeloutput. Tilstrækkelig dokumentation vil øge muligheden for, at en virksomhed kan genskabe tidligere versioner af en model. Det er derfor vigtigt, at den enkelte virksomhed har en politik for, hvordan den styrer og sikrer versionering af sine modeller.

Flere virksomheder har fremhævet, hvor væsentlig tilstrækkelig datakvalitet og datahåndtering generelt er for at sikre hensigtsmæssig brug af AI og for risikostyring. Det indebærer bl.a., at virksomheden har styr på de data, den bruger, hvor data kommer fra, hvordan data kan modificeres, og om den anvendte data kan føre til uhensigtsmæssige bias.

Virksomheden bør forholde sig til, om data indeholder følsomme oplysninger der ikke er nødvendige. Hvis en virksomhed bruger en model, der bygger på data, som virksomheden ikke forstår til fulde eller ved hvor stammer fra, kan modellen ende med at operationalisere og automatisere uønskede bias i data, som virksomheden egentlig ikke havde et ønske om at understøtte. Virksomheden kan derfor med fordel fokusere på indsamling og behandling af data, særligt hvis den pågældende model skal bruges på regulerede områder. Hvis virksomheden indhenter data fra tredjeparter, kan det være relevant at sikre, at data ikke ændrer karakter, da det kan ændre modellen og dens output mærkbart i forhold til det forventede.

Effektiv styring af data

Adgangen til meget store datamængder har skubbet til den stadigt mere udbredte brug af AI. Udviklingen har ikke alene øget brugen af eksterne datakilder, men kan også tilskynde til at arbejde med at samle internt data ét sted. Det kan give mening rent forretningsmæssigt, men det kan også medføre helt åbenlyse risici: Jo mere data, der er samlet ét sted, jo større vil konsekvenserne af et eventuelt datalæk være. Derudover bør virksomheden forholde sig til, om medarbejdernes adgang til f.eks. følsomme personoplysninger, der kan indgå i et samlet datasæt, er berettiget. Med udgangspunkt i de eksisterende krav til datastyring bør den enkelte virksomhed tage stilling til, hvordan den

styrer egne data effektivt, herunder hvordan virksomheden opbevarer og giver adgang til data. Dette er en forudsætning for, at virksomheden effektivt og sikkert kan samle, opbevare og benytte store mængder interne data til brug for AI.

Endeligt bør virksomhederne være opmærksomme på, at en automatisering af en proces ikke som udgangspunkt betyder, at kompetente medarbejdere på området bliver overflødige. En domæneekspert, der tidligere har udført en opgave, som en virksomhed nu ønsker at automatisere, har ofte den dybe indsigt i problemstillingen bag den opgave, som modellen alene er sat til at løse. Hvis en model af en eller anden årsag ikke længere præsterer, vil domæneeksperten ofte kunne bidrage til at finde årsagen og indgå i en gentræning eller efterfølgende validering af modellen.

God praksis for brug af kunstig intelligens er, at virksomheden:

- ud fra et risikoperspektiv, har forholdt sig til, om og hvordan dens modeller baseret på AI adskiller sig fra virksomhedens øvrige modeller
- med afsæt i modellens væsentlighed, har en plan for gentræning og regelmæssig validering af modeller, herunder eksternt indkøbte
- med afsæt i modellens væsentlighed, har forholdt sig til tilstrækkelighed af interne ressourcer, herunder til modelvalidering
- har en politik for, hvordan den styrer og sikrer forskellige versioner af sine modeller
- har fokus på effektiv datastyring, herunder opbevaring og adgang til data.

3. Forklarlighed

Flere virksomheder har gjort opmærksom på, at mere komplekse modeller ofte præsterer bedre end simplere og mere forklarlige modeller. For mange virksomheder er netop forklarlighed dog et centralt kriterie for valget af modelspecifikationer. Det kan derfor ofte være nødvendigt at foretage en afvejning mellem performance og forklarlighed, når en virksomhed udvikler en model og tager den i brug.

Forklarlighed er i denne sammenhæng ikke et præcist defineret begreb. Det dækker over, i hvor høj grad det er muligt at forklare et udfald af en model for en interessent – altså hvordan og hvorfor en model er nået til et konkret output. Forventningen til graden og formen af forklarlighed afhænger med denne tilgang helt af modtageren, herunder om modtageren er intern eller ekstern, og af hvor indgribende de beslutninger, der bliver truffet på baggrund af modellen, er.

For effektivt at bruge en model, skal en virksomhed både have kompetencer til at udvikle modellen, så den er brugbar og effektiv, og forståelse for efterfølgende at udnytte den bedst muligt til det tiltænkte formål. Tilstrækkelig internt rettet forklarlighed vil understøtte, at brugeren af modellen reelt forstår den og forstår de styrker og svagheder, der er knyttet til den.

Afvejning af performance og forklarlighed

De virksomheder, Finanstilsynet talte med, er generelt enige om, at forklarlighed er et vigtigt emne i forbindelse med brugen af AI. For nogle er det ligefrem det vigtigste. For andre er det vigtigste modellens evne til at præstere og levere resultater, hvorefter spørgsmålet om forklarlighed må behandles og løses tilstrækkeligt. Finanstilsynet anerkender, at der kan være eksempler, hvor denne afvejning ikke er nødvendig, hvis den konkrete model sikrer både høj performance og forklarlighed.

Finansielle institutioner, som ikke har direkte kunderelationer, vil møde en lavere grad af forventning til den eksternt rettede forklarlighed af deres modeller. Interne krav til forklarlighed kan fortsat være høje, selv om den direkte kunderelation ikke stiller krav om høj forklarlighed overfor kunden. Det kan f.eks. være investeringsselskaber, som driver en fond, der skal matche afkastet af et indeks. Omvendt vil selskaber, hvis handlinger direkte berører forbrugere, ganske givet møde en højere grad af forventning til eksternt rettet forklarlighed. Det kan f.eks. være forsikringsselskaber, som skal afgøre, om en kunde kan få lov til at tegne en konkret forsikring eller har ret til at få udbetalt en forsikringssum. Det kan også være selskaber, som godkender eller afslår låneansøgninger fra forbrugere. I disse tilfælde kan forventningen til eksternt rettet forklarlighed indebære, at forbrugere får at vide, hvordan en konklusion er nået, og hvor forklaringen ikke blot består af tekniske specifikationer.

I de tilfælde hvor finansielle virksomheder skal lave denne afvejning, kan det være afgørende, hvor indgribende de resultater, som en model leverer, er – f.eks. hvilken effekt såkaldte false positives ville have på kunder eller virksomheden. Et eksempel kan være en AI-model til transaktionsmonitorering hos en udbyder af betalingstjenester, som automatisk kan stoppe en transaktion, hvis der er en tilpas høj sandsynlighed for, at transaktionen er et forsøg på svindel. Hvis modellens vurdering viser sig at være forkert, kan kunden kontakte sin betalingstjenesteudbyder og få gennemført transaktionen eller låst sit kort op. Den negative påvirkning er i dette eksempel så lav, at ekstern forklarlighed ikke som udgangspunkt bør stå i vejen for at implementere den mest effektive model. Der kan i dette eksempel forsat være høje krav til intern forklarlighed. Modsat vil det være afgørende, at f.eks. et forsikringsselskab, der vurderer, om en kunde er berettiget til en erstatning, kan forklare, hvad der ligger til grund for afgørelsen – særligt hvis der er tale om et afslag.

Kompetencer til afvejning af forholdet mellem performance og forklarlighed

Det er afgørende, at den enkelte virksomhed forholder sig til spørgsmålet om forklarlighed. Virksomheden bør kunne dokumentere sin afvejning af forholdet mellem forklarlighed og performance for sine konkrete modeller, når det vurderes relevant.

Det er et gennemgående tema for de virksomheder, som Finanstilsynet har spurgt, om kravene til brugen af AI nødvendigvis skal være større end kravene til de processer hvor mennesker løser samme opgaver. Risikoen forbundet med at skalere automatiserede, og i varierende grad autonome, processer kan være af en anden karakter end den risiko, der er forbundet med, at individuelle eksperter løser opgaverne. Derfor kan der være behov for andre processer og kontroller end dem, der gælder for de individuelle eksperter. I den forbindelse er forklarlighed vigtig. Den ekspert, der tidligere har udført opgaven, har også kunnet forklare et konkret udfald; men når matematik og statistik erstatter eller supplerer en funktion, vil forventningen til forklaringen også blive en anden.

Afvejning mellem performance og forklarlighed kan være kompleks. Opgaven kræver en forståelse af det samlede brugsscenarie, bl.a. forretningsformålet, brugerne, teknologien og eventuelle lovkrav. Derfor bør virksomheden tage stilling til, hvordan den sikrer, at modellen fungerer efter hensigten, og hvordan virksomheden vil definere og måle dette. En virksomhed, der vil bruge AI til at løse en konkret udfordring, bør f.eks. overveje, om modellens definition af performance er hensigtsmæssig.

Flere virksomheder fremhæver, at hvis ansvaret for at lave afvejningen ikke er placeret, kan den uvægerligt tilfalde udvikleren af modellen, som ikke nødvendigvis har dyb indsigt i fagområdet og dets faldgruber. Udvikleren kan samtidig have svært ved at definere, hvad forklarlighed betyder for andre kundegrupper eller faggrupper end den, som udvikleren selv tilhører. En virksomhed kan derfor med fordel inddrage et tværgående team af fagfolk i

afvejningen mellem forskellige hensyn. Det kan f.eks. være eksperter med domænekendskab og tekniske færdigheder samt kompetencer indenfor jura, compliance og risikostyring. Virksomheden bør vurdere kompetencerne hos de medarbejdere, der foretager afvejningen mellem performance og forklarlighed.

Forståelse af modellens resultater eller bias

Forskellige brugsscenarier medfører forskellige risici og behov for forklarlighed afhængigt af, hvor kompleks en model er. Jo mere direkte indgribende eller på anden måde væsentlig et brugsscenarie er for virksomheden eller de berørte kunder, jo større er behovet for ekstern forklarlighed forventeligt. Hvad angår kundenære områder er det væsentligt at huske, at det er kunden, og ikke en modeludvikler eller domæneekspert, der skal kunne forstå modellens resultat. Det er ofte væsentligt for virksomheden selv at kunne forstå og forklare modellens resultater eller bias. Det gælder særligt i tilfælde, hvor modellens anvendelsesområde har betydelig indvirkning på enten virksomheden eller dens kunder, eller hvor modellens resultater strider mod økonomisk teori. Et eksempel på det sidste kan være, at en models resultat indikerer, at den mindste af to ellers næsten identiske lejligheder er mest værd, eller at positive stød til BNP forværrer aktiekursen for en ellers cyklisk aktie.

Vurdering af relevante brugsscenarier

En virksomhed bør vurdere hvert enkelt brugsscenarie som led i modeludviklingen med henblik på at afveje forholdet mellem performance og behov for forklarlighed, hvor det er relevant. Det kan samtidig styrke muligheden for, at virksomheden får identificeret og håndteret risici på et så tidligt stadie som muligt. Det kan også bidrage til, at virksomheden forbedrer sin overvågning af modellerne.

Dokumentation af valg og overvejelser om mitigerende tiltag

Afvejningen mellem forklarlighed og performance for en given model kan resultere i, at en virksomhed vælger en model med bedre performance på bekostning af forklarlighed. Det vil potentielt øge behovet for risikostyring, herunder monitorering og kontrol. Ved brug af komplekse modeller, der er svære at forklare, bør virksomheden generelt kunne retfærdiggøre sine valg, f.eks. ved at dokumentere sine overvejelser i den forbindelse. Virksomheden bør også overveje mitigerende tiltag og dokumentere disse overvejelser. Virksomheden kan på den måde underbygge sin vurdering af, at modellen performer som tiltænkt, på trods af at virksomheden ikke nødvendigvis til fulde kan forklare hvorfor.

Virksomheden bør derfor forholde sig til, om øget kompleksitet er nødvendig, og om modellens performance bliver forbedret af tiltag, der retfærdiggør, at modellens udfald samtidig bliver sværere at forstå og forklare. Hvis en mere

simpel model præsterer lige så godt, bør virksomheden dokumentere, hvorfor den har valgt den mere komplekse specifikation.

En række bredt anvendte modeller, såsom LIME og SHAP, kan hjælpe med at forklare, hvad der ligger til grund for en AI-models resultater. Det er dog værd at holde sig for øje, at de modeller, der understøtter forklarlighed, i sig selv er modeller bygget på antagelser og metodevalg. Det er derfor relevant at overveje, om disse modeller bør gennemgå samme procedure som virksomhedens øvrige modeller. Hvis virksomheden har svært ved at forklare, hvordan de modeller, den bruger til at forklare resultatet af en AI-model, virker, kan det være tegn på, at virksomheden ikke bruger den mest hensigtsmæssige forklaringsmodel i den pågældende situation. Det kan også være et mere grundlæggende tegn på, at virksomheden bør vælge AI-modeller, der som udgangspunkt er lettere at forklare.

God praksis for brug af kunstig intelligens er, at virksomheden:

- med afsæt i modellens væsentlighed, dokumenterer sine overvejelser om forklarlighed for konkrete modeller, herunder forholdet mellem forklarlighed og performance for konkrete modeller
- med afsæt i modellens væsentlighed, har vurderet, hvilke kompetencer der bør indgå i afvejning mellem performance og forklarlighed
- kan forstå og forklare, hvilke uhensigtsmæssige resultater eller bias modellen kan være forbundet med
- vurderer brugsscenarier som led i modeludviklingen for at klarlægge krav og behov for modellens forklarlighed
- ved brug af komplekse modeller, der er svære at forklare, og med afsæt i modellens væsentlighed, kan berettige og dokumentere sine modelvalg samt overvejer og dokumenterer mitigerende tiltag.