Vejledning: Tiltag til at sikre brugen af kunstig intelligens

Til danske virksomheder og myndigheder

JANUAR, 2020

ERHVERVSSTYRELSEN DIGITALISERINGSSTYRELSEN FE CENTER FOR CYBERSIKKERHED

Introduktion: Hvad skal jeg bruge denne vejledning til?

Til ledelsen

- Denne vejledning kan bruges til at skabe en grundlæggende forståelse for relevante informationssikkerhedsovervejelser ifm. udviklingen og brugen af kunstig intelligens (AI)
- Du kan lave et lyn-tjek af, hvorvidt denne vejledning er relevant for din organisation på næste side, og på s. 7-8 kan du læse mere om, hvordan brugen af AI potentielt kan ændre din organisations risiko

Til udviklere af kunstig intelligens

- Denne vejledning kan bruges til at stifte bekendtskab med AI-specifikke trusler og tiltag, der bør tænkes ind i modeludviklingen
- Du kan identificere de tiltag, der er relevante for netop din situation på s. 6, ligesom du på s. 9 kan læse om de fire nye angrebstyper, som organisationer eksponeres for ved at bruge AI

Til systemejere[1]

- Denne vejledning kan bruges i forlængelse af gængse standarder og vejledninger (som fx ISO 27001 eller sikkerdigital.dk) til at identificere sikkerhedstiltag, som er særligt vigtige ifm. brugen af AI
- Du kan læse om, hvordan denne vejledning passer ind i din organisations øvrige informationssikkerhedstiltag på s. 5, og på den følgende side identificere de tiltag der er relevante for netop jer

1. Til personer ansvarlige for AI-systemet (inkl. informationssikkerheden ifm. brug af systemet).
Note: Visse tiltag i denne vejledning kræver høj teknisk forståelse, og er primært rettet mod udviklere af kunstig intelligens

Denne vejledning er rettet mod virksomheder og myndigheder...

...som selv anvender kunstig intelligens (AI)

...som overvejer at anvende AI

...som anvender en AI-baseret service fra en leverandør

Lyn-tjek: Kan du svare 'Ja' til de følgende spørgsmål?

- Har I udarbejdet løsningsspecifikke risikovurderinger for alle de projekter, hvor I anvender AI?
- Har I i jeres risikovurderinger taget højde for nye typer af angreb rettet mod AI (fx forgiftning af data)?
- Har eventuelle leverandører af AI-løsninger været inddraget ifm. udarbejdelsen af risikovurderinger?
- Har I en form for kvalitetskontrol af output fra jeres AI-algoritmer?
- Har I en plan B i tilfælde af, at jeres AI-løsning pludselig ikke skulle være tilgængelig?

Hvis du har svaret 'Nej' til et eller flere af disse spørgsmål, kan du med fordel søge inspiration i denne vejledning

Indholdsfortegnelse

Formål, baggrund og metode for vejledningen

Formål: Sikre at danske virksomheder og myndigheder tager de fornødne forholdsregler, når de anvender kunstig intelligens

- I denne vejledning præsenteres 16 specifikke tiltag til, hvordan myndigheder og virksomheder kan mindske risikoen for cyberangreb forbundet med brug af kunstig intelligens
- Vejledningen fokuserer på nye sikkerhedstiltag samt traditionelle tiltag med særlige overvejelser ifm. brugen af kunstig intelligens. Den erstatter således ikke andre mere generelle informationssikkerhedsstandarder og -vejledninger (fx ISO 27001) men bør ses som en forlængelse heraf
- Tiltagene er identificeret på baggrund af:
  - Interviews med internationale og nationale eksperter og forskere
  - Interviews med internationale og nationale it-sikkerhedsudbydere
  - Interviews med danske virksomheder og myndigheder, som anvender kunstig intelligens
  - Research ifm. *Analyse af kunstig intelligens i et sikkerhedsperspektiv*[1]

Definition: Begrebet 'kunstig intelligens' inkluderer i analysen maskinlæring og defineres

"Kunstig intelligens systemer baseret på algoritmer—dvs. matematiske formler—der ved at analysere og finde mønstre i data kan identificere den mest hensigtsmæssige løsning. Langt de fleste systemer varetager specifikke opgaver på afgrænsede områder til fx kontrol, forudsigelse og vejledning. Teknologien kan udformes til at tilpasse sin adfærd ved at observere, hvordan omgivelserne påvirkes af tidligere handlinger"

National strategi for kunstig intelligens, EU-kommissionen og OECD

1. Analysen kan findes på sikkerdigital.dk Note: For yderligere definitioner og uddybning af metode se *Ordforklaringer* på s. 34. Kilde: BCG analyse

Baggrund: Øget anvendelse af og politisk fokus på AI og informationssikkerhed i Danmark

Vejledningen er udarbejdet i et samarbejde mellem Digitaliseringsstyrelsen, Erhvervsstyrelsen, Center for Cybersikkerhed, Globeteam og Boston Consulting Group og står i forlængelse af *National strategi for kunstig intelligens* samt *National strategi for cyber- og informationssikkerhed*, der har til formål at styrke hhv. brugen af kunstig intelligens og informationssikkerheden i Danmark.

Vejledningen adresserer overlappet mellem de to strategier gennem et øget fokus på informationssikkerhed for virksomheder og myndigheder, som anvender AI. For yderligere information vedrørende generel informationssikkerhed se fx

- Sikkerhedstjekket (ERST)
- Cyberforsvar der virker (CFCS & DIGST)
- Sikkerdigital.dk (ERST & DIGST)
- Kravkataloget (DIGST)
- Informationssikkerhed i leverandørforhold (CFCS & DIGST)

Vejledningen skal tænkes ind som supplement til det eksisterende arbejde med informationssikkerhed

Sikkerhedsframeworks

Standarder (eksempler)

ISO 27001

NIST

enisa – EUROPEAN UNION AGENCY FOR CYBERSECURITY

Vejledninger (eksempler)

sikker digital DK

FE CENTER FOR CYBERSIKKERHED

DIGITALISERINGSSTYRELSEN

CIS Center for Internet Security

Eksempler på tiltag

- Stærke passwords
- Malware-beskyttelse
- Specificering af roller og ansvar
- Awareness, uddannelse og træning
- Opdatering af programmer
- Kontrol med hardware og software
- Begrænsning af brugeradgange
- Monitorering af netværk

Denne vejledning fokuserer på problemstillinger specifikke ifm. brugen af kunstig intelligens og erstatter således ikke ovennævnte tiltag; tværtimod er ovennævnte tiltag nødvendige for at sikre den grundlæggende informationssikkerhed

Kilde: BCG analyse; Center for Internet Security; ISO 27001; NIST; Center for Cybersikkerhed

Læsevejledning: Hvilke tiltag er særligt relevante for min organisations brug af AI?

Udbredelse

	Tiltag		Side	Særligt relevant…
1	**Intern risikovurdering**	K	7	**…for alle - *fundamentet for alle øvrige tiltag***
2	Risikovurdering af leverandører	K	10	…for alle der anvender leverandør til udvikling eller hosting af AI-løsningen
3	Træning af medarbejdere	K	12	…for alle der bruger eller udvikler kunstig intelligens
4	Gendannelsesplan og backup-løsning	K	14	
5	Datakryptering	K	15	
6	Penetrationstest af systemer	K	17	
7	Styring af modeludvikling og -træning	AI	18	…for alle der udvikler kunstig intelligens
8	Sikring af modellens robusthed	AI	19	
9	Undgå læk af modelparametre og -beregninger	AI	21	…for udviklere, hvis AI-modellens resultater er synlige for kunder/borgere
10	Begrænsning af inputdata	AI	22	…hvis AI-modellen modtager inputdata direkte fra kunder/borgere
11	Overvågning af input og output	AI K	23	
12	Beredskabsøvelser	K	25	…hvis AI-løsningen er kritisk for organisationens virke
13	Benchmarking af model	AI K	26	…hvis output fra modellen er uigennemskueligt og ikke let kan verificeres
14	Træn på manipuleret (adversarial) data	AI	27	…for udviklere, hvis AI-modellen modtager inputdata direkte fra kunder/borgere, og hvor validiteten af input er svær at bekræfte
15	Modificering af inputdata	AI	29	
16	Øge validiteten af modeller med hårde labels1	AI	30	

1. Se definition på s. 34

AI *AI-specifikt sikkerhedstiltag*

K *Også relevant for købere af AI-løsning - ikke kun udviklere*

Note: Flere af de AI-specifikke tiltag relaterer sig til angreb, som hidtil kun i begrænset omfang er set i praksis

Risikovurderinger bygger et solidt fundament for sikker brug af AI

Generelt

En risikovurdering er en kortlægning og scoring af alle de risici, som et projekt medfører, samt en vurdering af hvilke organisatoriske og tekniske foranstaltninger der bør implementeres. Brugen af AI introducerer nye sårbarheder og øger antallet af angrebsflader, hvilket kan ændre organisationens overordnede risikoprofil, og som derfor bør afspejles i risikovurderinger af projekter med brug af AI.

Implementering

Risikovurderinger bør...

- ...udarbejdes førend projekter igangsættes og bør prioriteres på ledelsesniveau
- ...indarbejdes i AI-udviklingsfasen, således at evt. tiltag relaterende til *security-by-design* kan inkorporeres
- ...løbende opdateres for at reflektere ændringer i både trusselsbillede og anvendelse
- ...også udarbejdes for evt. leverandører af AI-udvikling eller -hosting (læse mere herom i tiltag 2)

Elementer

Generelt bør risikovurderinger inkludere tre trin (som også beskrevet i Digitaliseringsstyrelsens 'Vejledning i it-risikostyring og -vurdering'):

- Risikoidentifikation
- Risikoanalyse
- Risikoevaluering

Format

Mht. format adskiller risikovurdering af AI-applikationer sig ikke væsentligt fra risikovurdering af andre projekter med relevans for informationssikkerhed. Der kan således med god grund tages udgangspunkt i eksisterende vejledninger og skabeloner; på **sikkerdigital.dk** kan der findes værktøjer for både myndigheder og virksomheder. De særlige overvejelser ved brug af AI beskrives på næste side.

Brug af AI medfører en række specifikke overvejelser ifm. risikovurderingen

Risiko-identifikation

- Et første skridt i risikovurderingen er at identificere mulige risici. I identifikationen af risici er det vigtigt at overveje truslen fra hackere såvel som truslen fra interne medarbejdere i form af misbrug eller utilsigtede fejl. Samtidig bør man overveje sårbarheder og afhængigheder i forretningsprocesser og aktiver
- I forhold til at identificere cybertrusler ifm. brugen af AI er *STRIDE*[1] frameworket særligt relevant – men generelle kataloger (fx ISO, NIST) er også anvendelige. Brugen af AI giver desuden anledning til fire nye angrebstyper (læs beskrivelser på næste side)
- I forhold til sårbarheder vil brugen af AI ofte medføre udvidelser til den eksisterende teknologiske infrastruktur, fx adoption af cloudløsninger, hvilket øger eksponeringen og angrebsfladen

Risiko-analyse

- Dernæst kan man i en risikoanalyse måle størrelsen af de identificerede risici. Risikoanalysen vil variere afhængig af den specifikke anvendelse af AI
- Der bør tages højde for AI-løsningens vigtighed for organisationens drift, typen af data modellen baseres på (fx ift. GDPR) og typen af AI, der anvendes. I forhold til sidstnævnte er der fx en mindre angrebsflade ved en intern applikation (fx mailsortering) end en brugervendt applikation (fx estimering af forsikringspræmie)
- Ansattes kompetencer er en anden vigtig faktor; både i forhold til risikoen for fejl ved brug af AI men også i forhold til evnen til at opdage og respondere på eventuelle brud

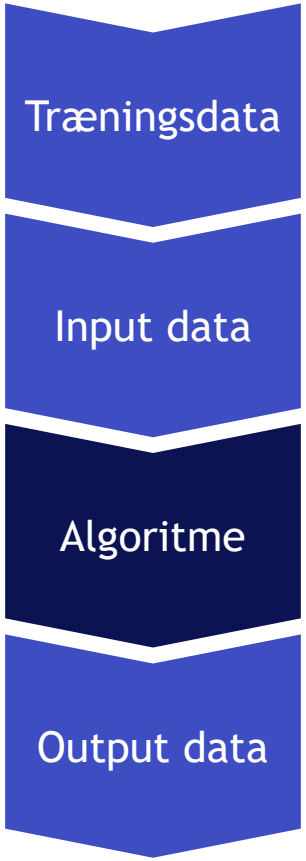
Risiko-evaluering

- Den hurtige udvikling på AI-området stiller endnu højere krav til organisationer ift. løbende at holde sig opdateret på nye typer af trusler eller tiltag og derved konstant holde risikovurderingen opdateret
- Fx kan der i forhold til AI være et særligt behov for at sætte nye kontroller ind i processen, da modellens *black box*-karakter stiller nye krav ift. kompleksiteten i at opdage brud
- Det, og en lang række andre tiltag til at forbedre informationssikkerheden ifm. brug af AI, kan du læse meget mere om på de følgende sider

1. Se fx 'The STRIDE Threat Model' fra Microsoft. Kilde: Digitaliseringsstyrelsen (2015): Vejledning i it-risikostyring og –vurdering; BCG analyse

Risikovurderingen skal tage højde for, at brugen af AI potentielt kan eksponere organisationen for fire nye typer af angreb

AI-proces



Angrebstype

Data poisoning (forgiftning af data) sker, når en aktør ændrer i træningsdata med formålet at manipulere output. Ændring i træningsdata kan både ske ved at ændre direkte i data (af interne eller eksterne aktører) eller, hvis algoritmen er selvlærende, ved at sende den en stor mængde inputdata af en bestemt art. Et kendt eksempel på sidstnævnte er Microsofts chatbot Tay, hvor Twitter-brugere på under 24 timer fik 'lært' Tay at skrive vulgært om en række ømtålelige emner.

Adversarial attacks (manipulering af data) sker, når hackere, typisk uden adgang til algoritmen, algoritmens parametre eller træningsdata, formår at 'snyde' modellen til at opnå en ønsket klassifikation ved at sende inputdata med en marginal ændring (fx ændre én pixel i et billede eller ændre bølgelængden på tale). Ofte kan ændringen ikke umiddelbart opdages af mennesker. Formålet kan bl.a. være *evasion*, hvor man ønsker at undgå en bestemt klassifikation, fx i en løsning hvor AI bruges til ansigtsgenkendelse.

Backdoor attacks (bagdørsangreb) sker ifm. AI typisk ved, at hackere eller interne får uretmæssig adgang til produktionskoden til algoritmen og ændrer koden (dvs. implementerer en *backdoor*) for at muliggøre et ønsket udfald ved en atypisk eller teoretisk umulig kombination af input variable. Det kan fx være input hvor *antal års joberfaring* > *alder*, som resulterer i en unormalt lav forsikringspræmie.

Data extraction (dataekstraktion) sker ved, at en aktør benytter information fra output til ét af to formål; 1) at bygge en surrogat/kopi model eller 2) at forstå bias i modellen. Ved førstnævnte kan den nye model fx sælges (som IP), bruges til at underbyde offeret på markedet, eller til at rekonstruere træningsdata. Ved sidstnævnte kan identifikationen af bias gøre det muligt at identificere kombinationer af inputdata, som resulterer i et ønsket udfald (også kaldet *impersonation*), hvilket bl.a. kan bruges i adversarial attacks.

Udbredelse

At udnytte de nye angrebsflader er **mere ressourcekrævende** for hackere end traditionelle angreb

Angrebene er **endnu ikke set i stor skala i praksis**, hvilket kan skyldes at de enten ikke er udført, opdaget eller offentliggjort

Basale sikkerhedstiltag mod traditionelle angreb bør som udgangspunkt være på plads før tiltag mod disse nye angrebsflader prioriteres

Kilde: BCG analyse; Patrick Hall: Proposals for model vulnerability and security; ekspertinterviews; nyhedsartikler

2. Risikovurdering af leverandører (K)

I. Risiko

Særligt relevant for: Alle som bruger en leverandør til udvikling/hosting af AI-løsningen

Overordnet risiko: Sikkerhedsbrud eller fejl hos leverandøren af AI-løsningen har negative konsekvenser for brugeren af AI-løsningen

- AI-løsninger kan udvikles og/eller drives eksternt (dvs. af leverandører), hvormed sikkerhedsbrud hos leverandøren kan påvirke brugeren af AI-løsningen negativt
- Fx kan en virksomhed/myndighed, der benytter tekstgenkendelse via en ekstern leverandør, blive udsat for læk af fortrolig data, hvis leverandørens AI-løsning udsættes for et angreb eller på anden måde kompromitteres
- Uden kendskab til leverandørens sikkerhedstiltag kan der være uoverensstemmelse mellem egen risikoprofil og leverandørens risikoprofil

II. Tiltag

Type af tiltag: Organisatorisk | Teknisk

Fokus for tiltag: Forebyggelse & beskyttelse | Opdagelse | Respons

Overordnet tiltag: Foretag risikovurderinger af leverandører mhp. at afklare risici ved brug af deres AI-løsning; herunder kortlægge leverandørens eksisterende sikkerhedstiltag

- Risikovurderingen kan tage form af et samarbejde mellem leverandør og klient, hvor implementeringen af særligt relevante sikkerhedstiltag kan præges af klienten
- Vurderingen kan bl.a. tage udgangspunkt i sikkerhedstiltagene i denne vejledning

Yderligere information

- Se fx Digitaliseringsstyrelsens 'Kravkatalog' (læs mere på næste side)
- Hvis AI-løsningen fører til brug af cloud, kan 'Vejledning i anvendelse af cloud-services' fra DIGST og CFCS med fordel konsulteres – især afsnit 3.4, 4 og 5

III. Effekt

Kvantitative og kvalitative effekter

- Tiltagets primære effekt er, at nødvendige sikkerhedsmæssige ændringer implementeres hos nuværende og fremtidige leverandører
- Tiltaget kan desuden benyttes som led i udvælgelsesprocessen af potentielle leverandører
- Derudover medfører tiltaget afklaring af leverandørers grad af informationssikkerhed, således at organisationens ønskede risikoprofil kan afstemmes på tværs af forsyningskæden

IV. Implementering

Deep dive

Beskrivelse af hvordan initiativet kan fungere i praksis

1. Kortlægning af leverandørens eksisterende og mulige fremtidige produkter og løsninger med brug af AI
2. Dialog med leverandør mhp. identifikation af særligt vigtige risici ved anvendelse af leverandørens AI-løsning
3. Afvejning af eventuelle risici forbundet med leverandørens AI-løsning, mhp. implementering af nødvendige sikkerhedstiltag for at opretholde organisationens ønskede grad af informationssikkerhed
4. Indarbejdelse af identificerede risikoelementer i leverandørkontrakt (se eksempler på næste side) eller adressering heraf i leverandørforholdet på anden måde

Forudsætninger for succesfuld implementering og mulige faldgruber

- Opmærksomhed på at leverandører kan tilbyde løsninger med AI, uden at dette fremgår klart; der kan i visse tilfælde spørges ind til dette, hvis løsningen fx bygger på store mængder data, kommer med anbefalinger osv.
- Risikovurderingen kan med fordel udarbejdes i fællesskab med leverandøren

Investeringer, løbende omkostninger og fremtidssikring af løsning

- Risikovurderingen foretages mindst én gang årligt eller ved eventuelle hændelser, samt hver gang en ny AI-leverandør inkluderes i organisationens forsyningskæde - i alle tilfælde med begrænsede omkostninger til følge

1. Se definition på s. 34

(K) *Også relevant for købere af AI-løsning – ikke kun udviklere*

Deep dive: AI-overvejelser kan indarbejdes direkte i kontraktkrav

2. Risikovurdering af leverandører

Ikke-udtømmende

Emne	Uddrag fra Digitaliseringsstyrelsens kravkatalog[1]	AI-specifikke overvejelser
Generelt	*"Leverandøren skal opdatere sin risikovurdering efter påbud fra Kunden om at inkludere en specifik trussel i risikovurderingen..." K2*	En opdateret risikovurdering bør også tage højde for de AI-specifikke trusler og tiltag beskrevet i denne vejledning
Træning	*"Leverandøren skal sikre, at Leverandørens medarbejdere og, hvor det er relevant, underleverandører ved hjælp af uddannelse og træning bevidstgøres om sikkerhed..." K16*	AI-specifikke trusler og sikkerhedstiltag bør indgå i uddannelse og træning; fx med udgangspunkt i emnerne nævnt i tiltag 3
Kryptografi	*"Leverandøren skal udarbejde og implementere en politik for anvendelse af kryptografi til beskyttelse af information i forbindelse med Kontraktens opfyldelse..." K43*	AI-modellers behov for løbende (gen)træning stiller krav til praksis for datakryptering, således at data er tilgængeligt for modellen når nødvendigt; tiltag 5 beskriver disse overvejelser
Backup	*"Leverandøren skal sikre, at der tages backupkopier af information, Programmel og system-images, der anvendes til Kontraktens opfyldelse..." K65*	Problematikken ved at lave backups af AI-modeller, som løbende lærer på nyt inputdata, gør dette til et vigtigt emne at diskutere med leverandøren; problematikken beskrives i tiltag 4
Udvikling	*"Leverandøren skal styre ændringer af Programmel indenfor udviklingslivscyklussen... ved hjælp af formelle procedurer for ændringsstyring..." K85*	*Black box* faktoren i AI-modeller gør det særligt relevant at sikre tilstrækkelig versionsstyring og dokumentation af modelindstillinger, træningsdata osv.; læs mere herom i tiltag 7
Testdata	*"Leverandøren skal omhyggeligt udvælge testdata og skal sikre, at testdata beskyttes og styres." K93*	Ifm. AI-udvikling vil der både være behov for test- og træningsdata, hvor splittet mellem de to og kvaliteten af data har stor indflydelse på modellens robusthed og derved sikkerhed - læs mere i tiltag 8
Lovkrav	*"Leverandøren skal sikre, at alle relevante lov-, myndigheds- og kontraktkrav, samt Leverandørens metode til overholdelse af disse krav, er klart identificeret, dokumenteret og opdateret..." K110*	På trods af at der ikke eksisterer AI-specifik lovgivning, vil der dog i mange tilfælde være behov for at kunne dokumentere årsagen bag en beslutning truffet af en AI, især ifm. beslutninger i det offentlige (læs mere på s. 31)

1. Digitaliseringsstyrelsens kravkatalog er som udgangspunkt udarbejdet til offentlige myndigheder, men er dog i mange sammenhænge (bl.a. her) også relevant for virksomheder. Kilde: Digitaliseringsstyrelsen (2017): Sådan stiller du krav til leverandører om informationssikkerhed - Katalog

3. Træning af medarbejdere (K)

I. Risiko

Særligt relevant for: Alle organisationer som anvender AI-løsninger, og træningen kan rettes mod både brugere, udviklere og driftsansvarlige

Overordnet risiko: Brugere, udviklere og driftsansvarlige begår fejl eller undervurderer risici grundet manglende forståelse for informationssikkerheden ifm. brug af AI

- Fejlhåndtering i udviklings- og brugsfasen kan resultere i, at den bagvedliggende data, model eller anden vigtig information uforsætligt lækkes
- Manglende forståelse for sårbarheder ved og trusler mod AI kan medvirke til, at organisationen tager utilstrækkelige sikkerhedsforanstaltninger, hvilket kan udnyttes af eksterne aktører – fx pga. manglende viden om nye angrebstyper
- Mangel på awareness og kompetencer kan desuden resultere i, at brud og læk opdages langsommere, hvilket kan forværre konsekvenserne heraf

II. Tiltag

Type af tiltag: Organisatorisk | Teknisk

Fokus for tiltag: Forebyggelse & beskyttelse | Opdagelse | Respons

Overordnet tiltag: Træn og undervis relevante medarbejdere i sårbarheder ved og trusler mod brugen af AI-baserede løsninger gennem e-learning moduler, on-site træning, introforløb etc.

- Træning bør tilpasses modtageren. Fx kan generel træning i sikker AI-brug tilbydes alle medarbejdere, sikkerhed ifm. udviklingsfasen kan rettes mod udviklere og information omkring overordnede risici kan rettes mod de driftsansvarlige
- Generelt bør der i træningen være fokus på den enkeltes ansvar (fx at rapportere mulige hændelser, sikre best practice datahåndtering, være opmærksom på relevante trusler, udføre kontrol af modeloutput mv.)

III. Effekt

Kvantitative og kvalitative effekter

- Træning reducerer risikoen for medarbejderfejl, gør det sværere for eksterne aktører at udnytte sårbarheder og understøtter beslutningsprocesser
- 60% af alle databrud hos adspurgte SME'er skyldtes uagtsomme medarbejdere eller leverandører ifølge Ponemon Institute (2018)
- ~40% af de i alt 5 mia. lækkede data i 2018 skyldtes ifølge Risk Based Security Inc. (2019) fejl forårsaget af medarbejdere
- CFCS og PET (2019) vurderer det sandsynligt, at ubevidste insidere kan være involveret i op mod halvdelen af alle registrerede sikkerheds-hændelser

IV. Implementering

Beskrivelse af hvordan initiativet kan fungere i praksis

1. Ledelsen skal prioritere træningen og afsætte de fornødne ressourcer
2. Kategorisér hvilken træning der er relevant for hvilke medarbejdere – se eksempler på næste side
3. Træning for udviklere og beslutningstagere bør foretages allerede inden AI-projektet igangsættes, så sikkerhed tænkes ind i løsningen og beslutningsprocessen
4. Evaluér effekten af træningen løbende, fx gennem beredskabsøvelser, og gentag processen efter behov

Forudsætninger for succesfuld implementering og mulige faldgruber

- Ledelsen skal bakke op om træningsinitiativet og dedikere tid og ressourcer hertil
- Træning skal ses som en løbende proces – det kan ikke blot implementeres og så glemmes

Investeringer, løbende omkostninger og fremtidssikring af løsning

- Omkostninger ifm. træningskurser etc. til medarbejdere nu og her. Løbende budget til onboarding af nye samt opfriskende træning af eksisterende medarbejdere
- Træningsmateriale skal løbende holdes opdateret, så det reflekterer det aktuelle sårbarheds- og trusselsbillede

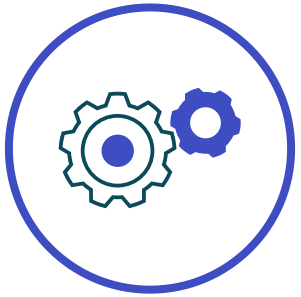
Kilde: Ponemon Institute (2018): 2018 State of Cybersecurity in Small & Medium Size Businesses; Risk Based Security (2019): The Year End 2018 Data Breach QuickView Report; CFCS og PET (2019): Cybertruslen fra bevidste og ubevidste insidere

(K) *Også relevant for købere af AI-løsning – ikke kun udviklere*

Deep dive: Træningen skal tilpasses medarbejdernes ansvar ifm. AI-løsningen

3. Træning af medarbejdere

Interne brugere



Udviklere

Driftsansvarlige[4]

	Interne brugere	Udviklere	Driftsansvarlige[4]
Gruppe	Brugere har den daglige interaktion med AI-løsningen, fx ved at anvende output der bruges til at understøtte beslutninger	Udviklerne er en del af AI-projektet fra start og har typisk ansvaret for vigtige dele af AI-løsningen, fx træningsdata og algoritme	Driftsansvarlige har ansvaret for informationssikkerheden ifm. AI-løsningen og den kontinuerlige drift i tilfælde af brud
Formål med træning	Understøtte sikker brug af AI-løsningen og gøre opmærksom på eventuelle faresignaler	Sikre at informationssikkerhed tænkes ind i AI-løsningen allerede i udviklingsfasen	Sørge for at brugen af AI er sikker og holdes ajour med viden om trusler, sårbarheder og juridiske forpligtelser
Relevante emner *Eksempler*	• Løbende kontrol af modeloutput • Basal viden om relevante trusler mod og sårbarheder ved AI-løsningen • Awareness ift. at identificere angreb samt rapportering af mulige brud • Test af brugerkompetencer • Information om dokumentationskrav ifm. fx GDPR (læs mere på sikkerdigital.dk[1] eller på Datatilsynets hjemmeside[2])	• Best practice ift. datahåndtering • Viden om trusler mod og sårbarheder ved anvendelsen af AI • Sikkerhedstiltag der er særligt relevante ifm. beskyttelse af AI (fx konvertering af inputdata, træning på manipuleret data) • Handlings- og gendannelsesplan i tilfælde af sikkerhedsbrud (læs mere på sikkerdigital.dk[3])	• Kendskab til relevante trusler mod og sårbarheder ved AI (se fx 'Analyse af kunstig intelligens i et sikkerheds-perspektiv'[5]) • Viden om best practice ift. governance af datastruktur og -håndtering, inkl. juridiske forpligtelser (fx GDPR) • Handlingsplan ifm. sikkerhedsbrud, inkl. gendannelsesplan • Vejledning i hvordan AI-aktivitet kan og bør overvåges

1. Sikkerdigital.dk: Databeskyttelse og GDPR; 2. Datatilsynet: Generelt om databeskyttelse; 3. Sikkerdigital.dk: Skabelon til It-beredskabsplan; 4. Inkluderer både IT-afdeling og evt. sikkerhedsafdeling; 5. DIGST, ERST og BCG (2020): Analyse af kunstig intelligens i et sikkerhedsperspektiv

4. Gendannelsesplan og backup-løsning (K)

I. Risiko

Særligt relevant for: Alle, dog især hvis AI-løsningen er kritisk for organisationens drift

Overordnet risiko: At AI-løsningen enten aldrig eller for langsomt bliver funktionsdygtig igen, hvis data eller kode gøres utilgængelig eller ødelægges

- Risikoen er også relevant for andre, generelle IT-løsninger, men øges fordi der typisk er et større databehov ved brug af kunstig intelligens
- Angreb kan introducere bias i AI-modellens output og kræve, at modellen gentrænes på 'rent' data – fx som følge af *data poisoning*[1] angreb mod træningsdata
- Dertil kan adgang til enten data eller selve AI-algoritmen reduceres eller helt ophøre, hvilket kan udfordre organisationens normale drift
- AI modeller med løbende læring fra fx brugerinteraktion (som besværliggør backup-processen), kombineret med AI's *black box* natur (som gør debugging praktisk talt umulig), øger konsekvenserne ved angreb der påvirker data eller modellen

II. Tiltag

Type af tiltag: Organisatorisk | Teknisk

Fokus for tiltag: Forebyggelse & beskyttelse | Opdagelse | Respons

Overordnet tiltag: Foretag jævnlige backups af kode og data, opbevar disse adskilt fra produktionsmiljøet og hav en plan for, hvordan modellen gendannes i tilfælde af brud

- Sikkerheden af backups skal være i fokus; fx opbevaring off-site, multifaktor-autentificering, adgangsstyring, krypteret indhold, forsinkelse på sletning af data, notifikation hvis kritiske handlinger udføres etc.
- Man bør overveje trade-off mellem backup-frekvens og lagerplads; høj frekvens stiller større krav til lagerplads men reducerer konsekvensen ved brud
- Bør udarbejdes i samspil med tiltag 7 om styring af modeludvikling og -træning

Yderligere information

- Se referencer på s. 35

III. Effekt

Kvantitative og kvalitative effekter

- Backup muliggør, at AI-modellen kan gendannes, hvortil en velfungerende gendannelsesplan understøtter, at dette gøres mest effektivt og med færrest mulige komplikationer
- Dertil tillader jævnlige backups identifikation af, hvornår problemerne opstod, og reducerer de negative konsekvenser ved at gå tilbage til en tidligere version
- Konsekvenserne ved manglende backups og dårlig respons- og gendannelsesplanlægning kan være store – fx er et *ransomware* angreb mod den amerikanske by Atlanta, der havde et betalings-krav på $52.000, estimeret til at ende med at koste omkring $17 mio. i alt (NY Times, 2019)

IV. Implementering

Beskrivelse af hvordan initiativet kan fungere i praksis

1. God praksis for opbevaring og backup af data og produktionskode opsættes – følg fx backup 3-2-1 reglen; 3 kopier af data og kode, 2 forskellige medier (harddisk, USB, cloud etc.), 1 off-site backup (fx i cloud)
2. Udførlig gendannelsesplan etableres og testes jævnligt. Planen bør beskrive hvordan driften opretholdes, hvis AI-løsningen er utilgængelig – se tiltag 12 om beredskabsøvelser for eksempler på tests heraf
3. I gendannelsesplanen skal der tages højde for AI-specifikke overvejelser, fx hvordan AI-modellen gentrænes, hvis der er behov for det, samt hvordan man efterfølgende tester, hvorvidt modelintegriteten er genetableret

Forudsætninger for succesfuld implementering og mulige faldgruber

- Gendannelsesplanen skal være specifik ift. processen, men samtidig favne over en bred vifte af risici. Den kan med fordel tage udgangspunkt i en kortlægning af, hvilke data og interne processer der er mest kritiske for AI-løsningen
- Kræver at beslutningstagere prioriterer tiltaget samt afsætter tid og midler til jævnlige tests

Investeringer, løbende omkostninger og fremtidssikring af løsning

- Omkostninger enten til opsætning af interne backupsystemer eller til tilkøb af ekstern lagerplads (fx cloud)
- Mindre omkostning til etablering af gendannelsesplan samt visse løbende udgifter til test og opdatering heraf

1. Se definition på s. 34. Kilde: The New York Times (2019): Ransomware Attacks Are Testing Resolve of Cities Across America

(K) *Også relevant for købere af AI-løsning – ikke kun udviklere*

5. Datakryptering (K)

I. Risiko

Særligt relevant for: Alle, men især AI-løsninger, hvor hele eller dele af modellen hostes eksternt, fx cloud, og hvis data er fortroligt

Overordnet risiko: Fortroligt data/persondata lækkes

- Internt i organisationen kan mangel på kryptering resultere i, at konsekvenserne ved brugerfejl, hvor data uforsætligt lækkes, øges
- Ydermere øger adgangen til rådata risikoen for, at interne medarbejdere kan misbruge deres rettigheder, da data kan stjæles, lækkes eller på anden måde udnyttes direkte, uden at medarbejdere også skal have adgang til fx krypteringsnøgle
- Udover interne processer er der også risiko for, at data lækkes eller misbruges, hvis denne deles uden for organisationen i rå form - fx hvis AI-løsningen hostes i cloud eller gennem anden ekstern leverandør

II. Tiltag

Deep dive

Type af tiltag: Organisatorisk | **Teknisk**

Fokus for tiltag: **Forebyggelse & beskyttelse** | Opdagelse | Respons

Overordnet tiltag: Kryptér data både ved opbevaring og i transit – fx til og fra cloud eller leverandør

- Dog er der et trade-off mellem sikkerhed og effektivitet i AI-udviklingsprocessen, da modellen ikke kan trænes på krypteret data; kræver at data fx dekrypteres én gang om måneden mhp. modeltræning, hvilket forsinker læringsprocessen
- Alternativt kan homomorfisk kryptering[1] eller hashing[1] anvendes, da begge metoder tillader at træne modellen og prædiktere outcome på det behandlede data; dog har disse også deres ulemper - læs mere om alle tre typer på næste side

Yderligere information

- Se referencer på s. 35

III. Effekt

Kvantitative og kvalitative effekter

- Kryptering af data minimerer konsekvenserne ved læk af data, da fortrolig eller værdifuld information i det lækkede data forbliver hemmelig, så længe data ikke kan dekrypteres – dog er det et trade-off, da data typisk skal dekrypteres førend det kan anvendes i modellen
- Brug af hashing er ligeledes forbundet med et trade-off, da det kan gøre det sværere at dokumentere beslutningsgrundlaget for en AI's afgørelse. Dette skyldes at data ikke umiddelbart kan tilbagekonverteres til standardformat

IV. Implementering

Beskrivelse af hvordan initiativet kan fungere i praksis

1. Tiltaget prioriteres og de nødvendige ressourcer afsættes i organisationen
2. Potentielle problemer forbundet med kryptering af data identificeres og tages med i den endelig beslutning om, hvorledes data bedst krypteres/hashes; fx om AI-modellen løbende skal træne, hvilket kryptering hindrer
   - Hvis data anvendes til træning eksternt (leverandør, cloud) kan dette fx sendes i hashet form
   - Hvis modellen trænes in-house, kan data eksempelvis dekrypteres én gang om måneden mhp. træning
3. For at styrke sikkerheden skal krypteret/hashet data holdes adskilt fra original data, krypteringsnøgle osv.

Forudsætninger for succesfuld implementering og mulige faldgruber

- Organisationen skal være konsekvent i kryptering af data og sikkerheden herom; fx bør data ikke opbevares i rå form, der bør være veldefinerede processer for anvendelse, opbevaring og adgang til krypteringsnøglen osv.
- Besværlighed ifm. modeludvikling og -brug kan afskrække nogle organisationer fra at implementere udførlig kryptering, og tiltaget bør derfor prioriteres af beslutningstagere og tildeles de nødvendige ressourcer

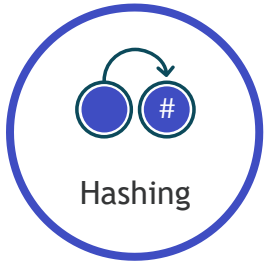
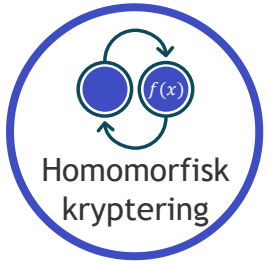
Investeringer, løbende omkostninger og fremtidssikring af løsning

- Mindre omkostning til køb/opsætning; løbende udgifter i form af tid og effektivitetsstab ifm. brugen af AI-løsningen

(K) *Også relevant for købere af AI-løsning – ikke kun udviklere*

Deep dive: Kryptering og hashing understøtter datafortrolighed og -integritet

5. Datakryptering

	Beskrivelse	Anvendelse
	Kryptering involverer konvertering af data til ulæselig kodetekst ved hjælp af matematiske beregninger og algoritmer. Gendannelse af meddelelsen kræver en tilsvarende dekrypteringsalgoritme og den originale krypteringsnøgle	• Formålet med kryptering er at sikre datafortrolighed under både opbevaring og transit til fx cloud eller ekstern leverandør • Kryptering bruges bl.a. til at beskytte filer lokalt eller i cloud, at sikre netværkskommunikation og at beskytte web- og mail-trafik • Kryptering er bredt anvendt og et standardtiltag ifm. data-håndtering og -beskyttelse
	Hashing foregår ved, at data af varierende længde konverteres til en værdi af fast længde – kaldet hashværdien. Mens kryptering er en tovejsfunktion, er hashing kun envejs, hvilket betyder at data ikke kan tilbagekonverteres til sit originale format	• Hashing er særligt relevant for AI-modeller med kategoriske variable, da det tillader at pseudonymisere data, inden det deles eksternt - uden at det påvirker modelbrugen, da AI-modellen stadig kan trænes og bruges på det hashede data • Juridisk kan hashing have nogle ulemper ift. kryptering, da hashet data ikke anses som fuldt anonymiseret, og derfor ift. GDPR stadig kan betragtes som persondata
	Homomorfisk kryptering er en form for kryptering, der tillader beregning på det krypterede data. Homomorfisk kryptering genererer et krypteret resultat, der, når det dekrypteres, matcher resultatet af beregninger, som om de var blevet udført på rådata	• Homomorfisk kryptering tillader, fra et AI-perspektiv, træning af AI-modellen og forudsigelse af outcome på krypteret data • På trods af at homomorfisk kryptering ikke er nyt, er det sjældent set anvendt i praksis, da det krypterede data er signifikant lang-sommere at udføre beregninger på – dog er der sket gennembrud de seneste år, og de første eksempler på kommerciel brug er set, fx til at anonymisere kræftpatienter

Kilde: EE Times; Hashed Out; Infosec Insights; Corti

6. Penetrationstest af systemer (K)

I. Risiko

Særligt relevant for: Alle brugere og leverandører af AI-løsninger

Overordnet risiko: Det pågældende system har uidentificerede sårbarheder

- AI-løsninger medfører ofte nye typer af datastrømme og -opbevaring, som ikke nødvendigvis er beskyttet af eksisterende sikkerhedsløsninger, hvorfor en penetrationstest efter implementeringen kan være hensigtsmæssig - fx efter indfasning af et cloud-baseret AI-system
- Der eksisterer desuden særlige risici for cyberangreb mod AI-løsninger, fx ifm. en ondsindet aktørs potentielle adgang til algoritmen eller algoritmens læringsdata, hvilket ikke nødvendigvis dækkes ifm. afviklingen af en mere generel penetrationstest af organisationens øvrige it-systemer

II. Tiltag

Type af tiltag: Organisatorisk | **Teknisk**

Fokus for tiltag: Forebyggelse & beskyttelse | Opdagelse | Respons

Overordnet tiltag: Hyr et it-sikkerhedsfirma til at forsøge at penetrere systemet, enten under kontrakt eller vha. såkaldt *bug bounty*, hvor der betales for identificerede sårbarheder

- Ved at foretage et kontrolleret cyberangreb mod sit eget system, kan man blotlægge uidentificerede sårbarheder, før disse udnyttes af en ondsindet aktør
- Konceptet betegnes også nogle gange som *white-hat hacking*[1] eller *ethical hacking*
- Særligt relevant i tilfælde, hvor et eksisterende system er blevet modificeret eller et nyt system er blevet introduceret, fx ved implementeringen af en AI-løsning
- Et alternativ er en sårbarhedsscanning, som typisk udføres 100% maskinelt; denne fokuserer dog på kendte trusler, og er således mindre relevant i en AI-sammenhæng

III. Effekt

Kvantitative og kvalitative effekter

- Tiltaget er bredt anvendt, og en veltilrettelagt penetrationstest kan bekræfte/afvise, at it-sikkerheden ved det pågældende system er konfigureret i overensstemmelse med *best practice*, fx om en given AI-løsning er tilstrækkelig sikret mod manipulation af læringsdata
- Identifikationen af sårbarheder kan også lede til opdagelsen af uidentificerede angreb mod organisationen, som tidligere har udnyttet disse sårbarheder

IV. Implementering

Beskrivelse af hvordan initiativet kan fungere i praksis

1. Et kvalificeret it-sikkerhedsfirma hyres, og der foretages en fælles afgrænsning samt eventuelt særligt fokus for penetrationstesten aftales, fx udefrakommendes adgang til algoritmen bag en AI-løsning
2. Penetrationstesten iværksættes, og der opretholdes løbende kontakt med det eksterne hackerhold for potentielt at ændre fokus for penetrationstesten løbende samt for at undgå unødvendig systemskade
3. En veldokumenteret sårbarhedsrapport afleveres af det eksterne hold
4. Der etableres nødvendige sikkerhedsforanstaltninger afhængigt af testens resultat

Forudsætninger for succesfuld implementering og mulige faldgruber

- Effektiviteten af tiltaget afhænger i høj grad af kvaliteten af de hyrede hackere, og der er derfor behov for at vurdere, hvilken leverandør der passer bedst til det specifikke formål (fx ift. AI-specifikke kompetencer)
- Kapacitet til at følge op på penetrationstestens resultater er nødvendigt

Investeringer, løbende omkostninger og fremtidssikring af løsning

- Omkostning ifm. hver enkelt penetrationstest; behøver dog ikke være omfattende, men kan tilpasses situationen

(K) *Også relevant for købere af AI-løsning - ikke kun udviklere*

7. Styring af modeludvikling og -træning AI

I. Risiko

Særligt relevant for: Alle udviklere ifm. modeludviklingsprocessen

Overordnet risiko: Integriteten af AI-løsningen kompromitteres

- I udviklingen, træningen, vedligeholdelsen og optimeringen af AI-løsningen er der risiko for, at integriteten af koden eller modellen kan blive kompromitteret af eksterne såvel som interne aktører, enten bevidst eller ubevidst
- I det omfang AI-løsningen er sat i drift og enten bidrager med beslutningsstøtte eller indsigter, er der en risiko for, at organisationen kan stå juridisk til ansvar for at dokumentere bevæggrundene for en given beslutning - og man bør derfor løbende have en indikation af, hvilke parametre der er udslagsgivende for outcome, hvilket data der er trænet på, og hvorledes dette har ændret sig over tid

II. Tiltag

Type af tiltag: Organisatorisk | Teknisk

Fokus for tiltag: Forebyggelse & beskyttelse | Opdagelse | Respons

Overordnet tiltag: Etablér en overordnet strategi for, hvordan modeludvikling og -træning skal foregå, og implementér et værktøj til versionsstyring

- Versionsstyringsværktøjer gør det bl.a. muligt at tracke og dokumentere ændringer i modelkoden, at flere kan arbejde på de samme filer samtidig, at forskellige versioner af kode kan sammenlignes, og at tidligere versioner af kode kan gendannes i tilfælde af fejl, brud på kodens integritet eller behov for dokumentation

Yderligere information

- Se referencer på s. 35

III. Effekt

Kvantitative og kvalitative effekter

- Hjælper udviklerne med at samarbejde og håndtere, at flere arbejder på den samme kode
- Øger sikkerheden ved at give kontrol over hvordan ændringer indføres, samt ved at distribuere kopier af koden, så den nemt kan gendannes i tilfælde af nedbrud, eller hvis enkelte systemer inficeres med ondsindet kode
- Giver øget ansvarlighed ved at gemme historik over, hvem der har indført hvilke ændringer i koden og hvornår

IV. Implementering

Beskrivelse af hvordan initiativet kan fungere i praksis

1. Adskilte it-miljøer til udvikling, test og produktion oprettes, og regler for flytning af kode herimellem etableres
2. Rammen for styring af modeludvikling og -træning fastlægges, herunder hvilket versionsstyringsværktøj der skal bruges, hvad det skal bruges til og af hvem; derudover foretages en risikovurdering af det valgte værktøj
3. Kode migreres til værktøjet – efter eventuelle forbehold er taget som følge af risikovurderingen
4. Der fastlægges klare regler for ændring af kode – se fx 'System change control procedures' i ISO 27002
5. Udviklere oplæres i at bruge værktøjet og følge de etablerede regler for kodeændring og -migrering

Forudsætninger for succesfuld implementering og mulige faldgruber

- Det er afgørende at alle udviklere oplæres i at anvende værktøjet, får tid til at vænne sig til det, og ikke bruger alternative systemer

Investeringer, løbende omkostninger og fremtidssikring af løsning

- Kræver ressourcer at implementere - både til opsætning og oplæring; værktøjer til versionsstyring findes både i gratis- og premium-udgaver
- Udviklere skal løbende holde sig opdaterede på bedste modeludviklingsmetoder og -løsninger på markedet

8. Sikring af modellens robusthed AI

I. Risiko

Særligt relevant for: Alle der udvikler AI-modeller, men med størst risiko for modeller baseret på små træningsdatasæt

Overordnet risiko: Modellen er sårbar overfor små ændringer i input og derved lettere at manipulere for eksterne, ondsindede aktører

- Hvis modeludvikling afviger fra *best practice*, øges risikoen for, at modellen opfanger forkerte strukturer i data. Fx lærte en model forskellen på græs og sne i stedet for hunde og ulve, da ulve havde sne i baggrunden og hunde græs
- Sådanne modeller vil ofte præstere godt på data, der ligner træningsdata, men have svært ved at generalisere - fx til billeder af ulve på græs. Af samme grund er modellen særligt sårbar overfor *adversarial attacks*[1], da få, simple ændringer i de dele af data, som modellen bruger til klassifikation, kan ændre resultatet markant

II. Tiltag

Type af tiltag: Organisatorisk | Teknisk

Fokus for tiltag: Forebyggelse & beskyttelse | Opdagelse | Respons

Overordnet tiltag: Hav fokus på robusthed under hele modeludviklingen, da det øger modellens evne til at håndtere ny, uset data, og dermed gør den sværere at manipulere gennem små ændringer i input (se eksempler på næste side)

- Generelt bør modeludvikling tænkes i et *security-by-design* perspektiv, hvor et af kernemålene for udviklingen og den løbende vedligeholdelse er, at adressere så mange sårbarheder som muligt - og derved besværliggøre angreb
- Modellen skal, som anden software, testes og igennem et Q/A-forløb udarbejdet af domæneeksperter, for at afklare bias og teste om data kan bære modelhypotesen

Yderligere information

- Se fx 'God praksis ved brug af superviseret machine learning' fra Finanstilsynet

III. Effekt

Kvantitative og kvalitative effekter

- Modelvalidering og -optimering øger modellens generaliserbarhed og styrker dermed robustheden mod fx *adversarial attacks*. Dette skyldes at en model, der træffer beslutninger på de korrekte informationer i data, også er sværere at snyde
- Derudover kan test af modellen på specifikke subgrupper, som er underrepræsenterede i data, øge chancen for at opdage input, som modellen er dårlig til at håndtere. Fx hvis der er få observationer fra ét specifikt geografisk område, som derfor systematisk fejlklassificeres af modellen

IV. Implementering

Deep dive

Beskrivelse af hvordan initiativet kan fungere i praksis

1. Tiltaget skal indarbejdes i interne processer allerede fra udviklingen af modellen påbegyndes
2. En modelejer med ansvar for hele modeludviklingen udvælges i organisationen, og kriterier til evaluering af modelperformance vælges, fx præcision, fejlrate etc. (se referencer på s. 35 for mere information)
3. Modellen trænes og testes gentagne gange, med fokus på at optimere modelperformance pba. af succeskriterierne sat op i trin 2. Det er vigtigt at testforløbet er veldokumenteret og foretaget pba. relevante cases
4. Modellen sættes først i drift, når den lever op til performancekravene

Forudsætninger for succesfuld implementering og mulige faldgruber

- Ledelsen skal afsætte ressourcer i form af tid og penge til, at AI-modellen kan udvikles korrekt fra start. Enten ved at opbygge de nødvendige kompetencer internt eller ved at købe rådgivning eller AI-løsningen hos en ekstern leverandør
- Modellen bør ikke sættes i produktion før den lever op til de ønskede standarder – fx ift. antallet af falske positiver

Investeringer, løbende omkostninger og fremtidssikring af løsning

- Kan kræve yderligere investering, hvis løsningen købes fra en ekstern leverandør, og alternativt kan der være omkostninger forbundet med at opbygge nødvendige kompetencer internt

1. Se definition på s. 34

Deep dive: Sikkerhed skal tænkes ind i hele AI-udviklingsprocessen

8. Sikring af modellens robusthed

Ikke udtømmende

AI-proces	Formål	Eksempler på implementering: Superviseret *Læring fra data med labels*	Eksempler på implementering: Usuperviseret *Læring fra data uden labels*
Dataindsamling og -forberedelse	**Kvaliteten og kvantiteten af data** er afgørende for, hvor mange og hvor gode eksempler modellen har at lære fra, hvilket påvirker modellens robusthed samt gør forgiftning af data sværere	Data skal valideres, tildeles korrekte labels og splittes i et fornuftigt test- og trænings-datasæt, og stabil adgang hertil skal sikres – særligt hvis modellen løbende gentrænes	Data skal valideres og derefter oversættes til sprog, som AI-modellen kan forstå, fx fra billeder til datasæt med pixels, højde, bredde etc.
Algoritme og træning	**Modelvalget** skal tilpasses formålet med AI-løsningen, da det øger modelperformance og dermed besværliggør udnyttelse for hackere. Det bør bl.a. overvejes, hvorvidt input er billeder, sekvenser (fx tekst eller musik) eller numeriske data, og om formålet er at estimere output præcist på observations- eller gruppeniveau	Valget af model skal passe til AI-løsningens formål og derudover balancere præcision og muligheden for fortolkning, fx kan flere forskellige AI-modeller bruges til reg-ression, hvor nogle er simplere end andre	Modellen bør, som for en superviseret AI, vælges pba. formålet og afveje indsigt og kompleksitet, da man også her har flere muligheder ifm. valget af AI-model
Evaluering og optimering	**Validering og optimering** af modellen skal foretages, så den performer så godt på nyt, uset data som muligt, da dette gør modellen mere robust	Tests skal foretages på data, som er holdt ude af træningsdata og derfor uset af mod-ellen, og krydsvalidering bør anvendes for at bruge den tilgængelige data mest effek-tivt. Dertil skal hyperparametre[1] optimeres, så modellen ikke *over-* eller *underfitter*	Sværere at evaluere pga. manglende labels. Man kan dog teste, hvor tydelig grupperingen er, ud fra hvor tæt data indenfor grupperne er samt hvor langt grupperne er fra hinanden (pba. såkaldt *Silhouette Coefficient*)
Modeltest og -brug	**Anvendelse af modellen eksternt** skal først ske, når modelperformance lever op til det ønskede niveau, så risikoen minimeres for, at en model der er under udvikling (og derved ikke færdigkalibreret) kan udnyttes	Modellen skal bl.a. have en tilpas høj præcision, få falske positiver eller bestå en penetrationstest, og dertil bør output kvalitetstjekkes løbende	Usuperviserede AI-modeller er særligt sårbare overfor *adversarial attacks*[1] og bør testes grundigt inden implementering - og løbende evalueres, mens de er i brug, fx vha. penetrationstest

1. Se definition på s. 34. Note: For yderligere referencer se s. 35. Kilde: Finanstilsynet (2019): God praksis ved brug af superviseret machine learning; BCG analyse

9. Undgå læk af modelparametre og -beregninger AI

I. Risiko

Særligt relevant for: Udviklere af AI-modeller, hvor modellens resultater er synlige for kunder/borgere, og i særdeleshed hvis data er fortroligt

Overordnet risiko: Ondsindede aktører kan bruge output fra AI-modellen til at kortlægge den underliggende data, genskabe algoritmen eller lære hvordan modellen kan snydes

- Hackere kan fx forsøge at afklare, hvorvidt enkelte individer er inkluderet i data, udfylde delvist kendte datasæt eller genskabe hele den bagvedliggende database
- Alternativt kan målet være at genskabe selve algoritmen, hvorefter denne enten kan udnyttes af hackere selv eller sælges, fx til én af offerets konkurrenter
- Sårbarheden opstår bl.a. når outputværdier offentliggøres, da det tillader hackere at teste effekten af små inputændringer på output. Derved kan det kortlægges, hvilke features der er vigtige, og hvornår ændringer i disse medfører skift i output

II. Tiltag

Type af tiltag: Organisatorisk | **Teknisk**

Fokus for tiltag: **Forebyggelse & beskyttelse** | Opdagelse | Respons

Overordnet tiltag: Offentliggør så lidt modeloutput og viden om modelparametre som muligt, da det gør det sværere for hackere at kortlægge den bagvedliggende data og algoritme

- Overvej fx forskellen mellem at vise den eksakte score for en AI-model der kredit-vurderer bankkunder, kontra hvis output begrænses til, om kunden har en 'høj' eller 'lav' score. Ved førstnævnte kan hackere, pba. små ændringer i input, nemmere kortlægge sammenhængen mellem in- og output og udnytte dette til egen fordel
- Ligeledes bør features, hyperparametre[1] osv. ikke være synlige for eksterne

Yderligere information

- Se referencer på s. 35

III. Effekt

Kvantitative og kvalitative effekter

- *Gradient masking*, hvor effekten af små inputændringer forsøges skjult, er en måde hvorpå output gøres sværere for hackere at bruge – dog kan effekten neutraliseres af sofistikerede hackere, hvis de alligevel formår at approksimere sammenhængen
- I tilfælde hvor modeloutput ikke kan/ønskes maskeret, kan der sættes en begrænsning for hvor hyppigt modellen kan anvendes (*rate limiting*)
- Shokri et al. (2016) viser, at man ved gentagne input til Googles og Amazons online AI-løsninger kan trække information om den underlæggende, personspecifikke data ud (*data extraction*[1])

IV. Implementering

Beskrivelse af hvordan initiativet kan fungere i praksis

1. Muligheder for at begrænse modeloutput identificeres, fx kan et it-sikkerhedsfirma hyres til at kortlægge, hvordan output fra AI-løsningen på nuværende tidspunkt kan udnyttes
2. Beslutningstagere vurderer, hvorvidt gevinsterne ved ændringerne vejer op for eventuelle negative effekter på fx brugeroplevelse og funktionalitet
3. Tiltaget implementeres – fx så modellen nu kun viser aggregeret output
4. Efter modellen er sat i drift, bør det løbende evalueres, hvilken information om modellen der bør være offentligt tilgængelig, og justeringer bør implementeres efter behov

Forudsætninger for succesfuld implementering og mulige faldgruber

- Forståelse af hvordan output i sidste ende skal anvendes ifm. brugen af AI-løsningen, så tiltaget ikke får utilsigtede negative implikationer for brugeroplevelsen eller funktionaliteten
- Manglende viden om hvordan output kan udnyttes, kan forårsage, at tiltaget ikke implementeres i tilstrækkelig grad

Investeringer, løbende omkostninger og fremtidssikring af løsning

- Kræver arbejdstid ifm. implementering og løbende evaluering

1. Se definition på s. 34. Kilde: Shokri et al. (2016): Membership Inference Attacks Against Machine Learning Models

10. Begrænsning af inputdata AI

I. Risiko

Særligt relevant for: Udviklere af AI-løsninger, hvor input kommer direkte fra kunder/borgere

Overordnet risiko: En ondsindet aktør afsender inputdata mhp. at fremprovokere særligt output eller at manipulere modellen

- Der kan bl.a. være tale om udnyttelse af særlige karakteristika ved datatypens mulige værdier (fx høje lydfrekvenser) eller logisk inkonsistent inputdata (fx ulige CPR-nr. for kvinder), i et forsøg på at manipulere modellen
- Et eksempel er såkaldte *dolphin attacks*, hvor en ondsindet aktør benytter ultralyd til at afgive skjulte kommandoer til en virtuel assistent med talegenkendelse - uden hverken ejerens godkendelse eller viden om hændelsen. Dette er bl.a. observeret i forbindelse med Amazons virtuelle assistent Alexa

II. Tiltag

Type af tiltag: Organisatorisk | **Teknisk**

Fokus for tiltag: **Forebyggelse & beskyttelse** | Opdagelse | Respons

Overordnet tiltag: Sæt begrænsninger på tilladt inputdata til modellen

- Begrænsningerne på inputdata kan fx opsættes ud fra en logisk betragtning om modellens anvendelsesområder, således at umuligt forekommende data ikke inkorporeres som input til modellen
- Det efterstræbes dermed at blokere for inputdata, som af egen kraft ikke er plausibelt (fx ved ikke-menneskelige lydfrekvenser ifm. talegenkendelse), eller som er i direkte uoverensstemmelse med modellens andre inputdata (fx hvis *antal års erfaring* > *alder*)

III. Effekt

Kvantitative og kvalitative effekter

- Begrænsning af inputdata medfører et mere velspecificeret anvendelsesområde, som besværliggør datamanipulation for en potentiel ondsindet aktør
- Endvidere vil særlige typer af cyberangreb kunne udelukkes, fx såkaldte *dolphin attacks*

IV. Implementering

Beskrivelse af hvordan initiativet kan fungere i praksis

1. Modellens typer af inputdata kortlægges, og risikoen for ondsindet udnyttelse af særlige karakteristika tilknyttet datatypen vurderes for hvert input
2. Der foretages en analyse af mulige restriktioner af inputdata, herunder bl.a. en analyse af datainputtets mulige værdier i et anvendelsesmæssigt perspektiv, fx en restriktion af intervallet for alder som input fra 18 til 85
3. Såfremt en eventuel datarestriktion vurderes irrelevant for modellens grundlæggende virke og har en veldefineret mængde af mulige værdier, implementeres restriktionen

Forudsætninger for succesfuld implementering og mulige faldgruber

- Det er afgørende, at begrænsning af inputdata er baseret på objektiv viden og har AI-løsningens formål in mente, så modellens effektivitet og brug ikke påvirkes unødigt

Investeringer, løbende omkostninger og fremtidssikring af løsning

- Implementering medfører en mindre stigning i omkostninger til udvikling

11. Overvågning af input og output AI K

I. Risiko

Særligt relevant for: Alle organisationer, dog særligt hvis input til AI-modellen kommer fra kunder/borgere, og hvis modellen bearbejder persondata eller fortroligt data

Overordnet risiko: Manipulation med inputdata kan udnyttes til at påvirke modeloutput eller opnå indsigt i den bagvedliggende data og algoritme

- Fx kan eksterne aktører anvende *data poisoning*[1] til at påvirke output af AI-modellen i en ønsket retning, ved at introducere fejlklassificeret data i træningen af modellen
- Derudover kan information om den underliggende data eller model kortlægges gennem *data extraction*[1] angreb, hvor modellen fødes store mængder inputdata
- Overordnet set bør modelinput og -output tjekkes systematisk for at reducere sandsynligheden for, at ondsindede eksterne aktører misbruger AI-modellen både under modeltræning og -brug

II. Tiltag

Type af tiltag: Organisatorisk | **Teknisk**

Fokus for tiltag: Forebyggelse & beskyttelse | **Opdagelse** | **Respons**

Overordnet tiltag: Tjek input til og output fra AI-modellen for anomalier - fx ændringer i frekvensen af specifikke udfald, distributionen af data eller antallet af forespørgsler

- For nyt inputdata kan distribution og karakteristika eksempelvis holdes op mod et 'rent' datasæt med kendte egenskaber - fx det nuværende træningsdata eller et andet renset datasæt, som kun bruges til validering af nyt data
- Dertil kan løsninger som fx NemID bruges til at sikre brugerintegritet
- Output bør ligeledes kvalitetstjekkes for at undersøge, om der er tegn på, at modellen er blevet kompromitteret - fx ved at sammenligne med historisk output

Yderligere information

- Se referencer på s. 35

III. Effekt

Kvantitative og kvalitative effekter

- Tiltaget handler på inputsiden om at reducere risikoen i modeltræning og på outputsiden om at reducere risikoen ved modelanvendelse
- Tjeks understøtter modelvaliditeten både hvad angår intern konsistens og tegn på misbrug i det enkelte datasæt, på tværs af datasæt og til validering af hele modellen
- Dertil øger løbende analyse af modelinput sandsynligheden for at identificere aktuelle angreb
- Selv små mængder korrupt data kan påvirke performance markant, fx som i Jagielski et al. (2018), hvor blot 8% korrupteret data ændrede outcome med over 75% for halvdelen af observationerne

IV. Implementering

Beskrivelse af hvordan initiativet kan fungere i praksis

1. Pba. en risikovurdering afgøres det, om tiltaget er værd at indføre. Fx er det særligt relevant for løsninger, hvor output vises til brugerne eller som løbende lærer, da disse er eksponerede for hhv. *data extraction* og *data poisoning*
2. Metode for datatjek etableres; se næste side for eksempler på, hvordan dette kan afhænge af trusselstypen
3. Data renses for de identificerede anomalier, og modellen gentrænes om nødvendigt. Dertil kan tjek af inputdata bidrage til løbende identifikation af cyberangreb - fx kan et unormalt højt antal forespørgsler være tegn på angreb
4. Tjek af input og output, samt hvordan data bevæger sig herimellem (flowet), udføres jævnligt; en audit af dataflow inkluderer fx at tjekke beregninger, sikre at observationer ikke slettes undervejs osv. (se referencer for mere info)

Forudsætninger for succesfuld implementering og mulige faldgruber

- Kræver regelmæssige opsyn og opfølgning på evt. anomalier
- Etableringen af et 'rent' valideringsdatasæt er kritisk, men kan være udfordrende, da det kræver indgående forståelse af mulige fejltyper, databegrænsninger etc.

Investeringer, løbende omkostninger og fremtidssikring af løsning

- Der skal dedikeres ressourcer til at udføre tiltaget, og definitionen af 'normalt' input/output skal holdes opdateret

1. Se definition på s. 34. Kilde: Jagielski et al. (2018): Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning

AI *AI-specifikt sikkerhedstiltag* K *Også relevant for købere af AI-løsning - ikke kun udviklere*

Deep dive: Overvågningen kan tilpasses til AI-løsningens brug

11. Overvågning af input og output

	Case-eksempel	Data	Eksempel på overvågning
	Forsikringsfirma der bruger AI-løsning til at foreslå brugerspecifik forsikring og dertilhørende forsikringspræmie • Overordnet trussel: Dataekstraktion • Eksempel på udnyttelse: En hacker kan sende en lang række forespørgsler, med små variationer i måden hvorpå forespørgselsformularen er udfyldt, til et forsikringsfirma. Derved kan han kortlægge sammenhængen mellem formular-input og forsikringspræmie, hvilket efterfølgende kan bruges til at genskabe den bagvedliggende model og/eller data • Andre eksempler: Investeringsfirma med AI-baseret investeringsmodel, offentlig institution der bruger AI til at fastsætte overførselsindkomst	Input: Kommer fra kunder i form af den indsendte information om alder, indkomst, tidligere ulykkes-historik, værdi af bolig-inventar osv. Output: Kun tilgængeligt for den enkelte kunde gennem en individualiseret forsikringspræmie	Analysér statistik omkring de foreslåede forsikrings-præmier, og om der er væsentlige ændringer mel-lem perioder (fx hvor mange estimater modellen udfærdiger eller gennemsnittet af de foreslåede forsikringspræmier). Dertil tjek for tegn på, at enkelte kunder misbruger modellen (fx ud fra IP-adresser) • Hvis der er risiko for dataekstraktion, kan fokus med fordel placeres på output. Således kan man opdage, hvis en kunde forsøger at trække store mængder information ud af modellen
	Søgemaskine der bruger AI-løsning til at danne autoudfyldninger i søgefeltet, pba. hvad tidligere brugere har søgt på • Overordnet trussel: Forgiftning af træningsdata • Eksempel på udnyttelse: En ondsindet virksomhed kan foretage et stort antal søgninger på konkurrentens navn sammen med fx ordene "snyder", "svindler" og "dårlig", for at påvirke hvad søgemaskinen vælger at autoudfylde og derigennem skade konkurrentens renommé • Andre eksempler: Virusprogram, spam-filter, chatbot	Input: Kommer fra brugere i form af alle ord og sætninger brugt i tidligere søgninger Output: Tilgængeligt for alle brugere igennem automatisk udfyldning af søgefeltet	Analysér statistik omkring søgninger (fx frekvensen af søgninger og specifikke ord), og om fordelingen ændrer sig væsentligt fra én periode til en anden. Dertil kan man tjekke, om søgningerne ofte stammer fra samme bruger. • Hvis der er risiko for forgiftning af trænings-data, herunder især hvis modellen lærer løbende, kan fokus med fordel placeres på input. Således kan man eksempelvis detektere, hvis der pludselig er et stort antal ekstra søgninger på ét specifikt ord

12. Beredskabsøvelser K

I. Risiko

Særligt relevant for: AI-løsninger som er kritiske for organisationens virke
Overordnet risiko: Enten at angreb mod AI-løsningen ikke opdages, eller at skaderne forbundet hermed ikke begrænses tidligt nok eller reduceres i omfang

- Angreb skal opdages, før de kan stoppes; eksempelvis kan manglende indblik i AI-specifikke angrebstyper resultere i, at angreb mod AI-løsningen går ubemærket hen – fx tager det, for sikkerhedsbrud generelt, i gnsn. 206 dage før de opdages (Ponemon, 2019) og i 41% af tilfældene opdages bruddet først af eksterne (Mandiant, 2019)
- Derudover kan der, når et angreb identificeres og skal afværges, være stor forskel på praktisk udførelse kontra den teoretisk funderede handlingsplan – fx kan responstiden stige grundet interne fejl eller vigtige procedurer ift. driften kan fejlagtigt blive stoppet

II. Tiltag

Type af tiltag: Organisatorisk | Teknisk
Fokus for tiltag: Forebyggelse & beskyttelse | Opdagelse | Respons

Overordnet tiltag: Foretag jævnlige beredskabsøvelser for informationssikkerheden, herunder effekten af angreb rettet mod AI-modellen – i stil med brandøvelser

- Øvelsen kan eksempelvis foregå ved, at et it-sikkerhedsfirma hyres til at simulere et angreb mod AI-modellen i form af ændret kode, forgiftning af data etc.
- Derudover bør øvelsen tage udgangspunkt i fastlagte planer, fx *business continuity*
- Dermed kan organisationen teste, hvorledes medarbejdere reagerer på en hændelse: Hvor hurtigt opdages bruddet? Bliver det afskærmet? Følger de ansatte de etablerede retningslinjer og handlingsplaner?

Yderligere information

- Se fx vejledninger/skabeloner til it-beredskabsplaner på Sikkerdigital.dk

III. Effekt

Kvantitative og kvalitative effekter

- Jævnlige beredskabsøvelser sikrer, at organisationen altid er forberedt på eventuelle brud; bl.a. fordi medarbejderne har klar forståelse for egne ansvarsområder og den generelle handlingsplan
- Dertil tillader beredskabsøvelser test af organisationens *business continuity*-plan, gendannelsesplan og backup-løsning for AI-modellen – se fx tiltag 4 i denne vejledning
- Intelligent, orkestreret hændelsesrespons kan øge responshastighed mod cyberangreb med 40x ifølge IBM

IV. Implementering

Beskrivelse af hvordan initiativet kan fungere i praksis

1. Beslutningstagere afsætter tid til, at øvelsen kan opsættes og udføres for relevante medarbejdere i organisationen
2. Øvelsen igangsættes, og der holdes nøje øje med, hvorledes medarbejderne reagerer, når deres vante arbejdsgang pludselig forstyrres af et sikkerhedsbrud – fx hvem der har ansvaret for hvad, og om procedurer følges
3. Dertil skal organisationen teste, hvordan driften opretholdes, hvis AI-løsningen på kort sigt er utilgængelig grundet fx ransomware angreb, eller hvis den er blevet manipuleret i et omfang, der sætter den ud af drift

Forudsætninger for succesfuld implementering og mulige faldgruber

- Forståelse for, at sikkerhed er en proces, som løbende skal testes, evalueres og tilpasses
- Øvelsen skal være realistisk, så medarbejderne træner deres reaktionsmønstre i en presset situation, og gøres på et tidspunkt og en måde, så den ikke forstyrrer vigtige elementer af den normale drift
- Generelt bør beredskabsøvelserne foretages jævnligt – fx som del af et sikkerhedsårshjul

Investeringer, løbende omkostninger og fremtidssikring af løsning

- Omkostninger ifm. nødvendige eksterne ydelser, fx *white hat hacking*[1] team, samt medarbejdertid ifm. øvelserne

1. Se definition på s. 34. Kilde: Ponemon (2019): Cost of a Data Breach - Report; Mandiant (2019): M-Trends 2019

K *Også relevant for købere af AI-løsning – ikke kun udviklere*

13. Benchmarking af model AI K

I. Risiko

Særligt relevant for: AI-løsninger hvor output ikke let kan verificeres, eller hvor der er stor risiko for interne fejl eller misbrug

Overordnet risiko: AI-løsningen manipuleres – bevidst eller ubevidst

- Hackere med adgang til selve koden kan foretage *backdoor attacks*[1]
- Alternativt kan *data poisoning*[1] bruges til at påvirke modellen indirekte, hvor manipuleret data langsomt inkluderes i træningssættet - fx så it-sikkerhedsløsninger vænnes til at misklassificere en malwaretype som godartet software
- Interne fejl såvel som bevidst internt misbrug kan ligeledes kompromittere integriteten af AI-applikationen – fx hvis forkert data fejlagtigt inkluderes i træningsdatasættet eller hvis udviklere misbruger deres adgangsrettigheder
- Generelt opstår der nye typer af manipulationsmuligheder grundet de potentielle vanskeligheder ved at kortlægge AI'ens beslutningsprocesser

II. Tiltag

Type af tiltag: Organisatorisk | **Teknisk**

Fokus for tiltag: Forebyggelse & beskyttelse | **Opdagelse** | **Respons**

Overordnet tiltag: Sammenlign outputtet af AI-løsningen med andre, validerede metoder; under træning og/eller løbende mens modellen er i brug

- Benchmarking kan ske mod 1) en tidligere version af AI-modellen, 2) en sideløbende AI-model trænet separat, eller 3) en simplere, mere transparent model (fx en lineær model)
- Store residualer mellem output af benchmarking- og produktionsmodellen kan være tegn på manipulation og bør undersøges nærmere

Yderligere information

- Se referencer på s. 35

III. Effekt

Kvantitative og kvalitative effekter

- Reducerer risikoen for, at manipulation går uset hen grundet manglende evne til at dokumentere beslutningsprocessen i AI-modellen
- Benchmarking af modellen understøtter desuden, at der er konsistens på tværs af modelversion, datasæt og modeltype, hvilket styrker model-validiteten og -robustheden. Eksempelvis kan et ens output på tværs af forskellige AI-algoritmer indikere, at inputtet er validt
- Reducerer risikoen ifm. fejl og misbrug af medarbejdere, da fejlagtig brug kan identificeres og misbrug besværliggøres med den ekstra kontrol

IV. Implementering

Beskrivelse af hvordan initiativet kan fungere i praksis

1. Det skal vurderes, hvorvidt det er nødvendigt at foretage benchmarking løbende, mens modellen er i brug, eller blot ifm. udvikling med enkelte stikprøver i brugsfasen; bør fx foregå sideløbende med anvendelsen, hvis output fra AI-modellen er svært at verificere - eksempelvis hvis AI-modellen bruges til automatisering af beslutninger
2. Den optimale metode til benchmarking identificeres; fx kan to ens AI-modeller, som dog trænes på forskellige datasæt, bruges til at validere hinanden (vha. residualer) og dermed afsløre tegn på manipulation
3. Benchmarking samt proces for løbende validering implementeres

Forudsætninger for succesfuld implementering og mulige faldgruber

- Der skal benchmarkes mod en metode, som repræsenterer ikke-manipuleret output, og det er derfor vigtigt, at benchmarket ikke selv er kompromitteret
- Valget af benchmarkingmodel skal tilpasses AI-løsningens formål, eksempelvis kan man vælge en simplere regressionsmodel, hvis AI-løsningen bruges til en form for regression

Investeringer, løbende omkostninger og fremtidssikring af løsning

- Kræver tid og kompetencer - især til opsætning, men også til løbende tjek

1. Se definition på s. 34

AI *AI-specifikt sikkerhedstiltag* K *Også relevant for købere af AI-løsning – ikke kun udviklere*

14. Træn på manipuleret (adversarial) data AI

I. Risiko

Deep dive

Særligt relevant for: Udviklere af AI-modeller, hvor input kommer direkte fra kunder/borgere, og især hvis modellen bruges til beslutningsstøtte

Overordnet risiko: AI-model snydes til at give et bestemt output

- AI-modeller baseret på neurale netværk bliver stadig mere udbredte, og de bruges især når datamængden er stor og kompleks, fx når data er tekst, billede og lyd
- Neurale netværk har vist sig at være sårbare overfor *adversarial attacks*[1], hvor manipulerede datainputs, der ikke umiddelbart kan skelnes fra almindeligt data, bliver klassificeret forkert (se eksempler på næste side)
- *Adversarial attacks* er dog til dato sjældent set i praksis
- Forkert klassifikation kan have stor betydning, hvis klassifikationen bruges til beslutningsstøtte, såsom hvis klassifikationen af en ansøgning, et røntgenbillede eller en samtale bruges som støtte i en beslutning om medicinsk behandling

II. Tiltag

Type af tiltag: Governance | **Teknologisk**

Fokus for tiltag: **Forebyggelse & beskyttelse** | Opdagelse | Respons

Overordnet tiltag: Inkludér eksempler på *adversarial* data i det datasæt, som AI-modellen trænes på, da det øger modellens robusthed overfor denne type af angreb

- Logikken bag tiltaget er, at man selv genererer *adversarial* data – men associeret med korrekte labels - som AI-modellen derfor kan lære at genkende
- I udarbejdelsen af eksemplerne på *adversarial* data, bør man kortlægge den inputtype eller -kombination, hvor AI-modellen er særligt sårbar overfor et *adversarial attack*, og derefter inkludere repræsentative eksempler herpå

Yderligere information

- Se referencer på s. 35

III. Effekt

Kvantitative og kvalitative effekter:

- Tiltaget er til dato hovedsageligt testet i en teoretisk og akademisk kontekst, ligesom truslen om *adversarial attacks* også sjældent er set i praksis; effekten af tiltaget er derfor også svær at kvalificere
- Adskillige forskere har i forsøg formået kraftigt at reducere raten af forkerte klassifikationer vha. metoden, fx reducerede Goodfellow et al. (2015) raten fra 89% til 18% - og der kommer løbende nye akademiske artikler om emnet, som beskriver nye variationer af både angreb og forsvar

IV. Implementering

Beskrivelse af hvordan initiativet kan fungere i praksis

1. Ledelsen allokerer tid for udviklerne til at implementere tiltaget
2. Udviklerne genererer deres eget *adversarial* data baseret på det data, som modellen i forvejen er trænet på, eksempelvis ved at downloade et værktøj eller bibliotek, som kan lave små ændringer i datasættet
3. Modellen gentrænes med inklusion af det genererede datasæt
4. Tiltaget forankres som en tilbagevendende opgave (vedligehold)

Forudsætninger for succesfuld implementering og mulige faldgruber

- Succes påkræver at ledelsen afsætter ressourcer til tiltaget, at udviklere med høje tekniske kompetencer er til rådighed, samt at tiltaget løbende vedligeholdes
- Det er svært/umuligt at inkludere alle typer af (ny) *adversarial* data i sin modeltræning, hvorfor implementering bør kombineres med løbende evaluering og gentræning

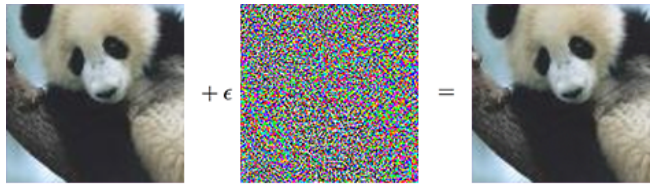
Investeringer, løbende omkostninger og fremtidssikring af løsning

- Det er nødvendigt, at udviklere løbende dedikerer tid til vedligehold af tiltaget med seneste udvikling indenfor *adversarial attacks*

1. Se definition på s. 34. Kilde: Goodfellow et al. (2015): Explaining and Harnessing Adversarial Examples

Deep dive: Træning på manipuleret data er relevant for flere datatyper

14. Træn på manipuleret (adversarial) data

Type	Beskrivelse af *adversarial attacks*[1]	Eksempel
Billeder	Ved at tilføje en smule specielt udvalgt 'støj', som er usynligt for det menneskelige øje, kan AI-algoritmen fx snydes til at tro, at et billede af en panda med overvejende sandsynlighed (99.3%) er en abe	
Lyd	*Adversarial attacks* mod lyd ses udført ved at tilføje lyd til det originale lydinput, det kan fx være subtile jingler, tilfældig støj eller lyd i frekvenser mennesker ikke kan høre (et såkaldt *dolphin attack*). Det er typisk sværere rent teknisk at beskytte sig mod angreb med hørbar lyd, som ligger indenfor det almindelige lydbillede og frekvenserne i det originale lydinput	
Tekst	I tekstanalyse kan *adversarial attacks* fx resultere i ændret klassifikation af sætninger. Som i eksemplet til højre, hvor skift af enkelte ord medfører, at teksten går fra at være klassificeret som *fake news* til *real news*	"trump supporter whose brutal beating by black mob was caught on video asks: *what happened to america?*" → "trump supporter whose ferocious beating by black gangsta was caught on tape demands: *what happened to america?*" 97% fake news 100% real news

1. Se definition på s. 34. Kilde: Carlini & Wagner (2018): Audio Adversarial Examples: Targeted Attacks on Speech-to-Text; Goodfellow, Schlens & Szegedy (2014): Explaining and Harnessing Adversarial Examples; Kuleshov, Thakoor, Lau & Ermon (2018): Adversarial Examples for Natural Language Classification Problems

15. Modificering af inputdata AI

I. Risiko

Særligt relevant for: Udviklere af AI-løsninger, hvor validiteten af input er svær at bekræfte, fx i billeder hvor visse ændringer ikke kan identificeres af mennesker

Overordnet risiko: AI-model snydes til at give et bestemt output, som følge af små ændringer i inputdata

- Den overordnede risiko stammer fra truslen om *adversarial attacks*[1], som tiltaget *Træn på manipuleret (adversarial) data* også adresserer (se s. 27)
- Modeller der tager rå inputdata er nemmere at udnytte, da det giver ondsindede aktører større kontrol ifm. manipulation af inputdata – fx ændring af specifikke pixels i billeder, så indholdet fejlklassificeres, bevidste stavefejl i tekst der ændrer stemningen (*sentiment)* i udsagnet, eller ved at tilføje små forstyrrelser til lydfiler, så indholdet transskriberes forkert

II. Tiltag

Type af tiltag: Organisatorisk | **Teknisk**

Fokus for tiltag: **Forebyggelse & beskyttelse** | Opdagelse | Respons

Overordnet tiltag: Modificér inputdata, så der tilføjes små ændringer eller tilfældigheder til data, da det gør det mere besværligt for hackere nøjagtigt at manipulere output til egen fordel

- Billeder kan eksempelvis komprimeres fra PNG til JPEG, tilføjes tilfældig støj (fx pba. en statistisk fordeling) eller udglattes – dog skal modifikationen være 'svag' nok til, at inputdata fortsat kan fortolkes og bruges af modellen
- Logikken er, at en ondsindet aktør mister en del af kontrollen med sine *adverserial samples*, hvorved det bliver sværere at manipulere modellen i en bestemt retning

Yderligere information

- Se referencer på s. 35

III. Effekt

Kvantitative og kvalitative effekter

- Das et al. (2018a) formår med JPEG-kompression at eliminere op til 98% af angreb mod billeder, udført med nogle af de stærkeste angrebsformer - fx *Carlini-Wagner Attacks*
- Ligeledes formår Das et al. (2018b) at reducere succesraten for angreb mod lydinput fra 92% til 0% vha. MP3-kompression og rateændringer
- Udover billedgenkendelse og lyd, kan modificering af inputdata også reducere truslen fra angreb mod tekst. Pruthi et al. (2019) viser, at man ved at bytte om på bogstaver, droppe endelser etc. i tekststrenge kan reducere effekten af bevidste stavefejl i klassificeringen af filmanmeldelser

IV. Implementering

Beskrivelse af hvordan initiativet kan fungere i praksis

1. Muligheden for og graden af datamodificering vurderes – fx bør trade-off mellem modificering af data og bevaring af original fortolkning overvejes ift. effekt på modelperformance kontra sikkerhed
2. Dernæst skal den optimale modificeringsmetode identificeres (fx PNG til JPEG eller ved at tilføje tilfældig støj)
3. Tiltaget implementeres, og modellens performance testes; implementering bør ikke påvirke den generelle modelpræcision og -ydeevne unødvendigt
4. Effekten af modificeringen bør løbende evalueres

Forudsætninger for succesfuld implementering og mulige faldgruber

- Teknisk know-how i forbindelse med valg af den optimale modificeringsmetode
- Der kan være en risiko for, at tiltaget mister effekt, hvis modificeringen er for svag - men omvendt er der risiko for, at modelperformance kompromitteres ved for kraftig modificering

Investeringer, løbende omkostninger og fremtidssikring af løsning

- Omkostninger ifm. udarbejdelse af den tekniske løsning og implementering. Dertil løbende omkostninger til regelmæssigt at kortlægge de optimale metoder og opdatere løsningen, så den følger med udviklingen i trusselsbilledet

1. Se definition på s. 34. Kilde: Das et al. (2018a): Shield: Fast, Practical Defense and Vaccination for Deep Learning; Das et al. (2018b): ADAGIO: Interactive Experimentation with Adversarial Attack and Defense for Audio; Pruthi et al. (2019): Combating Adversarial Misspellings with Robust Word Recognition

16. Øge validiteten af modeller med hårde labels AI

I. Risiko

Særligt relevant for: Udviklere af modeller trænet på data med hårde labels[1], og hvor der er en reel trussel om adversarial attacks

Overordnet risiko: AI-model snydes til at give et bestemt output

- Den overordnede risiko stammer fra truslen om *adversarial attacks*[1], som tiltaget *Træn på manipuleret (adversarial) data* også adresserer (se s. 27)
- Metoderne beskrevet i dette tiltag reducerer risikoen forbundet med *adversarial attacks* mod AI-løsninger med 'hårde labels', hvilket vil sige labels hvor medlemskab i en kategori er binært tildelte (fx ja/nej, 0/1 eller kat/hund/hest)
- Hårde labels gør modellen sårbar, hvis angribere kan identificere de specifikke ændringer der skal til for at kategoriseringen ændres (fx fra 0 til 1)
- Truslen er tidligere set ifm. spamfiltre, men brugen af AI kan øge omfanget samt åbne op for nye sårbarheder

II. Tiltag

Type af tiltag: Organisatorisk | **Teknisk**

Fokus for tiltag: **Forebyggelse & beskyttelse** | Opdagelse | Respons

Overordnet tiltag: Anvend *label smoothing*[1], *ensemble modeller*[1] og/eller *defensiv destillering med to modeller*[1]

- Ved *label smoothing* udglattes hårde labels, så modellen er mere moderat - fx skift 0 til 0.2 og 1 til 0.8, så outcome i stedet er lav/høj sandsynlighed og ikke binært
- I *ensemble modeller* trænes flere modeller, hvorefter output vægtes til ét samlet resultat; baseret på idéen, at det er sværere at snyde flere modeller end én
- *Destillering* fungerer ved, at man træner to modeller i stedet for én. Den første producerer bløde labels, hvilket den sekundære (produktions-)model trænes på

Yderligere information

- Se referencer på s. 35 og yderligere forklaring af de tre modeltyper på s. 34

III. Effekt

Kvantitative og kvalitative effekter

- Alle de tre nævnte metoder søger at gøre AI-modellen mere moderat i sit output, hvilket øger performance og generaliserbarhed, og dermed robustheden overfor *adversarial attacks*
- Goibert & Dohmatob (2019) finder at *label smoothing* øger robusthed overfor *adversarial attacks*
- Tramèr et al. (2018) byggede en *ensemble model*, der vandt første runde af NIPS 2017 ved at opdage ~95% af *adversarial attacks*
- Papernot et al. (2016) reducerede succesraten for *adversarial attacks* med 90% med *defensiv destillering*

IV. Implementering

Beskrivelse af hvordan initiativet kan fungere i praksis

1. Gevinsterne ved implementering skal vejes op mod omkostningerne herved - fx kræver *label smoothing* generelt mindre computerkraft end *ensemble modeller*
2. En eller flere af metoderne implementeres som del af AI-modellens udvikling
3. Test af modelrobusthed overfor *adversarial attacks* bør være del af modeloptimeringen

Forudsætninger for succesfuld implementering og mulige faldgruber

- Løsningen skal indtænkes allerede i udviklingsfasen af AI-modellen (*security-by-design*)
- Det er vigtigt at bemærke, at tiltaget ikke eliminerer risikoen for, at inputdata kan manipuleres for at opnå et bestemt output - men at det derimod besværliggør udnyttelse, og dermed reducerer risikoen herfor

Investeringer, løbende omkostninger og fremtidssikring af løsning

- Begrænsede omkostninger i forbindelse med udvikling og eksekvering; dog kan implementering stille større krav til computerkraft og forsinke udviklingen
- Implementeres én gang og opdateres automatisk ifm. senere gentræning

1. Se definition på s. 34. Kilde: Goibert & Dohmatob (2019): Adversarial Robustness via Label-Smoothing; Tramèr et al. (2018): Ensemble Adversarial Training: Attacks and Defenses; Papernot et al. (2016): Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks

Nuværende lovgivning stiller ikke specifikke krav til informationssikkerheden ifm. brugen af AI – men mangel på handling kan dog give juridiske problemer

Simplificeret overblik

Konklusion

- Der eksisterer ingen AI-specifik lovgivning til dato, men en række initiativer herom er sat i værk både på nationalt og EU-retligt plan
- Særligt tre faktorer er relevante ifm. brugen af AI fra et juridisk synspunkt; 1) at AI ofte indbefatter store mængder (potentielt følsomt) data, 2) at AI i visse situationer selv er i stand til at træffe beslutninger, og 3) at AI ofte anses som en *black box*
- Eksisterende lovgivning omkring databehandling og dokumentationskrav er således meget relevant ifm. brugen af AI lige såvel som visse sektor-specifikke reguleringer
- Ift. informationssikkerhed gælder således samme praksis for AI som ellers; selv uden lovgivning på området, kan mangel på handling skabe juridiske problemer (fx med hensyn til GDPR, Forvaltningsloven, almindelige forvaltningsretlige principper samt produktansvar og -sikkerhed)

Eksempler på relevant lovgivning

Myndigheder

- **Forvaltningsloven** beskriver partsrettigheder (borger og offentlig forvaltning) ved sagsbehandling, fx relevant når myndigheder træffer afgørelser baseret på AI
- **Offentlighedsloven** beskriver krav til myndigheders transparens ifm. aktindsigt i dokumenter, der er indgået til eller oprettet af en myndighed i relation til bl.a. anvendelsen af AI

Begge

- **Databeskyttelsesforordningen (GDPR) og databeskyttelsesloven** beskriver fysiske personers rettigheder ifm. både myndigheder og virksomheders behandling af personoplysninger, inkl. ved 'automatiske afgørelser'. Eksempelvis ved profilering med AI, hvor den registrerede har visse særlige rettigheder. Derudover kan der henvises til Datatilsynets vejledninger for bl.a. risikovurdering, konsekvensanalyse og behandlingssikkerhed[1]

Virksomheder

- **Produktsikkerhedsloven** regulerer de sikkerhedskrav, som er stillet til selve produktet. Brug af AI i interne processer er således ikke direkte dækket heraf. Er produktet kommet til veje vha. AI, kan dette dog stadig være relevant, hvis produktet eller produktbrugeren har forvoldt en skade
- **Diverse sektorspecifikke reguleringer** kan ligeledes have en hvis relevans ved brugen af AI, fx loven om ligebehandling på forsikringsområdet, hvor en AI ikke må forskelsbehandle bl.a. mænd og kvinder ved estimering af forsikringspræmier

1. Find dem på Datatilsynets hjemmeside under *Generelt om* databeskyttelse. Kilde: Konsultation med Kromann Reumert; Datatilsynet

Metode: Tiltagene i denne vejledning er identificeret vha. interviews med eksperter, udbydere, myndigheder og virksomheder, samt fra faglitteratur

Metode

Tiltagene præsenteret i denne vejledning er udvalgt på baggrund af en større, generel kortlægning af risici forbundet med brugen af kunstig intelligens for danske virksomheder og myndigheder.

Kortlægning af risici: Overordnet er AI-baserede løsninger eksponeret for tre trusler; læk af data, ændring af data/algoritme og ændring af tilgængelighed. Truslerne udføres af én af tre aktører; eksterne hackere, ondsindede medarbejdere (misbrug) eller ubevidste medarbejdere (fejl). De ni risikoscenariers sandsynlighed og konsekvens for den enkelte organisation afhænger af en række faktorer; vigtigheden af AI-løsningen for driften, typen af data benyttet, graden af tillid fra kunder/borgere, populationsstørrelsen (kunder/borgere) samt organisationens sårbarheder ifm. teknologisk infrastruktur, adgange og rettigheder samt adfærd og retningslinjer.

Udvælgelse af tiltag: Baseret på ovenstående er der identificeret 34 sikkerhedstiltag, som alle adresserer én eller flere af de ni trusler, og dækker både organisatoriske og tekniske tiltag. Tiltagene er identificeret pba. interviews med eksperter, udbydere, myndigheder og virksomheder samt research af faglitteratur. De 16 tiltag i denne vejledning fokuserer på nye sikkerhedstiltag samt traditionelle tiltag med særlige overvejelser ifm. brugen af AI, og dækker alle tre faser af it-forsvaret; forebyggelse & beskyttelse, opdagelse og respons. Vejledningen erstatter således ikke andre mere generelle informationssikkerhedsvejledninger men er specifikt rettet mod styrkelse af sikkerheden ved brugen af AI.

Bidragsydere

Nedenstående institutioner har bidraget til vejledningen gennem interviews og feedback (se flere næste side)[1]

1. Visse bidragsydere er udeladt efter eget ønske. Kilde: BCG analyse

En lang række brugere af kunstig intelligens har bidraget til vejledningen

Private virksomheder

TDC Group	Bruger AI i flere funktioner, bl.a. kommercielt (fx til at fremskrive markedstendenser), kundeservice (fx chat- og emailbots), streaming (fx anbefalinger) og generelt (fx automatisering)
PFA PENSION	Automatiserer processer, personaliserer tilbud til kunder og yder beslutningsstøtte - bruger bl.a. sprogteknologi til at gruppere forespørgsler fra kunder og foreslå svar
LEO Innovation Lab	Bruger maskinlæring til at optimere og generere nye indsigter i kliniske studier samt som værktøj i en platform, hvor brugere kan tage billeder af hudsygdomme og få en diagnose
TRADESHIFT	Bruger maskinlæring til automatisk at indlæse og kode fakturaer samt at berige og berigtige fejlbehæftede data hos kunder, når kunderne skal onboarde nye leverandører
boost·ai	Udvikler virtuelle agenter og 'konversationel' AI, der eksempelvis bruges i chatbots
2021.AI	Er en teknologivirksomhed, der leverer en AI platform, Grace, til udvikling og drift af AI modeller - med AI Governance modulet sikres, at disse lever op til etiske og regulatoriske krav
Forecast	Bruger maskinlæring til at 'læse' projektbeskrivelser mhp. at automatisere og optimere projektstyringen, fx via øget transparens om tidsforbrug og optimeret ressourceallokering
BOTSUPPLY	Udvikler sit eget framework til at analysere sprog med maskinlæring (natural language processing), der består af adskillige moduler, og som bl.a. bruges i chatbots og voicebots
cashbackAPP	Udvikler loyalitetsprogram direkte til kundens eksisterende betalingskort, som inkluderer maskinlæring, der automatisk kan anbefale den enkelte bruger relevante butikker i nærheden
RegTech PASS muinmos	Udvikler software til finansielle institutioner, som automatisk klassificerer deres kunder og sikrer mod fejlsalg ved at automatisere hele den lovgivningsmæssige compliance-proces
corli	Tilbyder en machine learning platform til diagnostisk beslutningsstøtte i kritisk dialogbaseret triagering indenfor sundhedssektoren
bizmiq	Udvikler bl.a. chatbots og herunder et AI-værktøj, der gennem dialogen med en chatbot kan finde frem til persontype ud fra teksten
whaii	Automatiserer virksomheders screeningsproces ifm. rekruttering og matcher kandidaters værdier, personlighed, kulturelle præferencer, faglige evner mv. med arbejdsgiveren
Aguardio	Samler data om vandforbrug for private kunder og hoteller via IoT-løsning og eksperimenterer med at bruge maskinlæring til at foreslå vandbesparende adfærd

Offentlige myndigheder

UDVIKLINGS OG FORENKLINGS STYRELSEN	Bruger maskinlæring til adskillige formål, herunder automatisering af processer samt scoring af indberetninger og identifikation af outliers med henblik på at identificere muligt snyd
Miljø- og Fødevareministeriet Landbrugsstyrelsen	Bruger maskinlæring til at analysere satellitfotos af markarealer for at monitorere landbrugsaktiviteter
Styrelsen for Arbejdsmarked og Rekruttering	Bruger maskinlæring til at forudsige udviklingen i efterspørgsel på arbejdsmarkedet samt i datamining af jobopslag (fx hvilke kompetencer der efterspørges)
	Tester flere løsninger, hvor maskinlæring anvendes til risikobaseret udvælgelse af, hvor tilsyn skal foretages
atp=	Anvender løsninger baseret på maskinlæring i flere dele af organisationen, bl.a. bruges modeller til clustering af data med henblik på at finde lighedstræk mellem sager om snyd
KØBENHAVNS KOMMUNE	Bruger og tester adskillige løsninger, bl.a. sortering af telefonkø, fordeling af post, prioritering af henvendelser, matching af borgere med aktiviteter og detektion af snyd
norddjurs kommune	Fordeler og journaliserer indgående post automatisk ved at bruge maskinlæring til at klassificere posten
midt regionmidtjylland	Bruger maskinlæring til at analysere sundhedsdata fra adskillige kilder og forudsige hvilke patienter, som vil blive akut indlagt indenfor det næste år mhp. at kunne målrette hjælp
REGION Region Hovedstaden	Tester AI-løsning, hvor IoT-devices trådløst måler og sender data om postoperative patienter til en computer, som vha. maskinlæring analyserer data og varsler evt. komplikationer
Ankestyrelsen	Bruger en algoritme til at sortere den indkomne post til relevante teammapper i sin administration
NÆVNENES HUS	Tester koncept, hvor AI bruges til at aflæse indkomne klagesager og automatisk fremsøge lignende sager og afgørelser
ERHVERVSSTYRELSEN	Anvender maskinlæring til kontrol og tilsyn på blandt andet regnskabsområdet og i forbindelse med virksomhedsregistreringer

Note: Visse bidragsydere er udeladt efter eget ønske. Kilde: Interviews med virksomheder og myndigheder

Ordforklaringer: Centrale begreber og termer

Adversarial attack: AI-specifik angrebstype. Når hackere formår at 'snyde' modellen til at opnå en ønsket klassifikation ved at sende inputdata med marginale ændringer (fx af få pixels i et billede eller bølgelængden på tale), som ofte ikke umiddelbart kan opdages af mennesker

Benchmark model: En kendt model, hvor output kan anvendes til at sammenligne og validere, om resultater fra AI-modellen agerer som forventet – fx kan afvigelser mellem de to modeller indikere misbrug eller fejl

Cyberangreb: Cyberangreb er hændelser, hvor en aktør forsøger at forstyrre eller få uautoriseret adgang til data, systemer, digitale netværk eller digitale tjenester. En aktør skal i denne sammenhæng forstås bredt og dækker både over hacking af eksterne aktører samt tilsigtet misbrug af interne aktører

Data extraction: AI-specifik angrebstype. Sker ved, at en aktør udnytter output til ét af to formål; 1) at bygge en surrogat/kopi model eller 2) at forstå modelbias. Ved 1) kan den nye model fx sælges (som IP), bruges til at underbyde offeret på markedet eller til at rekonstruere træningsdata. Ved 2) kan identifikationen af bias udnyttes til at identificere input, som resulterer i et ønsket udfald, og derved anvendes i *adversarial attacks*

Data poisoning: AI-specifik angrebstype. En aktør ændrer træningsdata med formålet at påvirke output. Ændring kan både ske direkte i træningssættet (af interne eller hackere med adgang) eller, hvis modellen er selvlærende, ved at sende en stor mængde inputdata af en bestemt art

Defensiv destillering: Forsvarsmetode for AI-modeller, der bruges til klassificeringsproblemer, hvor to modeller trænes i stedet for én. Den første trænes på 'hårde' labels (se beskrivelse nedenfor), for at bibeholde maksimal præcision, hvorefter outputtet, der er sandsynligheder, bruges som 'bløde' labels i træningen af den anden model, som så sættes i produktion og bruges til klassificeringen

Ensemble model: En model, hvor output fra en række forskellige AI-modeller aggregeres til ét samlet output, som derefter anvendes – fx kan typen af algoritme eller træningsdatasættet variere på tværs af modellerne

Robusthed: AI-modellens evne til at finde sammenhænge, som kan generaliseres på tværs af observationer og datasæt. Fx testet ved *out-of-sample* performance for at sikre, at modellen ikke forklarer hhv. for meget eller for lidt af variationen i træningsdata (dvs. *over*- eller *underfitter* træningsdata)

Input: Data som benyttes i en AI-model, der er sat i produktion. Input er de uafhængige variable (fx alder, indkomst, tekst), som en trænet AI-model konverterer til output; input skal derfor ikke forveksles med træningsdata, der er input til AI-modellen i træningsfasen

Hashing: Transformering af data til en fast længde, kendt som hash-værdien. Kendetegnet ved, at et givent input altid giver den samme hashværdi, og at denne ikke kan tilbagekonverteres

Homomorfisk kryptering: En krypteringsmåde, hvor strukturen i rådata bevares under krypteringen, så matematiske operationer foretaget på det krypterede data tilsvarer de samme operationer foretaget på rådata

Hyperparameter: Parameter i AI-modellen, hvis værdi er fastlagt inden læringsprocessen begynder, fx læringshastighed el. kompleksitet, modsat andre parametre, hvis værdier afledes under modeltræningen

Hårde labels: Opdeling hvor klassificeringen af, hvorvidt en observation tilhører en gruppe eller ej, er binær (fx billedet forestiller en hund: ja; en kat: nej). I modsætning til bløde labels, hvor der angives en sandsynlighed for at observationen tilhører den pågældende gruppe (billedet forestiller med 87% sandsynlighed en hund og med 2% sandsynlighed en kat)

Informationssikkerhed: Bred betegnelse for de samlede foranstaltninger til at sikre informationer i forhold til fortrolighed, integritet (ændring af data) og tilgængelighed. I arbejdet indgår blandt andet organisering af sikkerhedsarbejdet, påvirkning af adfærd, processer for behandling af data, styring af leverandører samt tekniske sikringsforanstaltninger

It-sikkerhed: Dækker over informationssikkerhed for data, der behandles i it-systemer. It-systemer omfatter hardware og software og kan være både uafhængige eller forbundet med andre systemer i et netværk

Krydsvalidering: Bruges til at evaluere AI-modellen og optimere hyperparametre ved at dele træningsdata i to (eller flere) 'folde' og derefter bruge én fold til validering og de resterende til træning af modellen (også kendt som udviklingsdatasættet). Herefter skiftes hvilken fold, der bruges til validering, og modellen gentrænes indtil alle folde har været brugt til validering (mens resten bruges til træning) - heraf 'krydset'

Kryptering: Kodeteknik der får information til at fremstå uforståelig for tredjepart og på den måde hemmeligholdt. Kryptering bruges ofte til at sikre information, der skal sendes via ikke-sikre kommunikationskanaler som f.eks. internettet.

Label smoothing: Forsvarsmetode til AI-modeller med kategoriske labels, hvor labels udglattes for at forhindre modellen i at forudsige disse for selvsikkert under træning og derved have lav robusthed – fx ved at ændre klassificeringsmålet fra 1 til 0,9 og fra 0 til 0,1

Output/outcome: Resultatet af AI-modellen; fx sandsynlighed for at tilhøre et givent label, den prædikterede værdi pba. en regression eller en opdeling af data i grupper (clusters)

Phishing: Forsøg på at manipulere en person til, i god tro, at videregive personlige oplysninger eller klikke på inficerede filer eller links til falske hjemmesider

Produktionskode: Skal forstås bredt og inkluderer al kode relevant for AI-løsningens udvikling og drift; både kode fra tidligere stadier i modeludviklingsprocessen såvel som koden bag den endelige, funktionsdygtige AI-løsning - til fx datahåndtering (ETL-koden), den endelige AI-algoritme, behandling af modeloutput etc.

Security-by-design: Tilgang til modeludvikling, hvor sikkerhed er tænkt ind i hele udviklingsprocessen – fx ved at have fokus på at minimere modellens sårbarheder gennem multifaktorautentificering, *best-practice* kode, løbende penetrationstests etc.

Træningsdata: Datasættet anvendt til modeltræning; normalt opdeles data i træningsdata, som fx kan bruges til krydsvalidering, og testdata, som bruges til den endelige validering af modelperformance

Backdoor attacks: AI-specifik angrebstype. Sker ifm. AI typisk ved, at hackere eller interne får uretmæssig adgang til algoritmen og ændrer koden (dvs. implementerer en backdoor) for at muliggøre et ønsket udfald ved en atypisk eller teoretisk umulig kombination af input variable. Det kunne fx være input hvor *antal års joberfaring* > *alder*, som kunne resultere i en unormalt lav forsikringspræmie

White hat hacking: Når it-sikkerhedsfirmaer, ofte hyret af organisationen selv, forsøger at hacke en organisation eller en specifik it-løsning (fx en AI-model) for at identificere sårbarheder i informationssikkerheden, så disse kan adresseres inden, ondsindede aktører udnytter dem

Kilde: Center for Cybersikkerheds *Ordforklaringer*; projektets egne definitioner

Bilag: Yderligere referencer for mere information

Tiltag		Referencer
1	Intern risikovurdering	• 'Threat Modeling AI/ML Systems and Dependencies' - Marshall, Parikh, Kiciman og Kumar (2019, Microsoft)
4	Gendannelsesplan og backup-løsning	• 'Intro to data management - Security & Backup' - Carnegie Mellon University • 'Saving a machine learning Model' – GeeksforGeeks
5	Datakryptering	• 'Hashing vs Encryption - The Big Players of the Cyber Security World' - Infosec Insights (2019) • 'Encrypt your Machine Learning' – Medium/Corti (2018) • 'These Three Security Trends Are Key to Decentralize Artificial Intelligence' – Hackernoon (2018)
7	Styring af modeludvikling og -træning	• For eksempler på *Version Control Systems* se fx DVC (data) og Git (kode) – begge er *open source* • For yderligere overblik søg fx efter *MLOps* eller *Version control ML models*
8	Sikring af modellens robusthed	• 'Metrics to Evaluate your Machine Learning Algorithm' - Towards Data Science (2018)
9	Undgå læk af modelparametre og -beregninger	• 'Privacy Attacks on Machine Learning Models' – InfoQ (2019) • 'How to attack Machine Learning (Evasion, Poisoning, Inference, Trojans, Backdoors)' - Towards Data Science (2019)
11	Overvågning af input og output	• For eksempel på et *valideringssystem*, se fx Tensorflow Data Validation • 'Auditing Artificial Intelligence' – ISACA (2018)
13	Benchmarking af model	• 'Proposals for model vulnerability and security' – O'reilly (2019)
	Træn på manipuleret (adversarial) data	• 'Review of Artificial Intelligence Adversarial Attack and Defense Technologies' – Qiu, Liu, Zhou og Wu (2019) • 'Explaining and Harnessing Adversarial Examples' – Goodfellow, Shlens and Szegedy (2015) • 'Towards Robust and Verified AI: Specification Testing, Robust Training, and Formal Verification' – DeepMind (2019)
15	Modificering af inputdata	• 'Evasion attacks on Machine Learning (or "Adversarial Examples")' - Towards Data Science (2019)
	Øge validiteten af modeller med hårde labels	• 'Adversarial Examples: Attacks and Defenses for Deep Learning' – Yuan, He, Zhu og Li (2019):

Formålet med listen er at give mulighed for at læse mere om teknikken bag specifikke tiltag, men den har ikke til hensigt at fremhæve specifikke hjemmesider eller forfattere. Der tages generelt ikke ansvar for indhold og eventuelle synspunkter i referencerne angivet på siden her.

Bilag: Udtømmende liste med it-sikkerhedstiltag relevante ifm. brugen af AI

I denne vejledning er kun beskrevet tiltag, som er nye, eller som i væsentlig grad ændrer karakter ifm. brugen af AI.
Beskrivelser af alle tiltag kan findes i *Analyse af kunstig intelligens i et sikkerhedsperspektiv*

	Forebyggelse & beskyttelse *Før angrebet sker*	Opdagelse *Når angrebet sker*	Respons *Efter angrebet er sket*
Organisatorisk	• **Risikostyringens ramme** defineres, herunder bl.a. afklaring af acceptabel risiko, allokering af ressourcer, fastlæggelse af metode til at måle risici mv. • **Risikovurdering** foretages for hvert (AI) projekt for at kortlægge og tydeliggøre de risici, som organisationen er udsat for • **Retningslinjer og politikker** udformes for organisationen for grundig og løbende monitorering og styring af risici • **Roller og ansvar** fordeles, så dedikerede kompetencer afsættes, og ansvar tydeligt kan henføres til enkeltpersoner på ledelses-/bestyrelsesniveau		
	• Træning af medarbejdere • Awareness-kampagne • Test af brugerkompetencer • Positivliste af software og devices • Risikovurdering af leverandører • Styring af brugere og adgange • Styring af opdateringer og vedligehold • Styring af modeludvikling og -træning AI	• Træning af medarbejdere • Awareness-kampagne	• Træning af medarbejdere • Handlings- og kommunikationsplan • *Business continuity*-plan • Gendannelsesplan • Beredskabsøvelse • Feedback-loop af *lessons learned*
Teknisk	• Penetrationstest af systemer • Malwarebeskyttelse • Datakryptering • Etablering af netværksgrænser og barrierer • Sikring af modellens robusthed AI • Undgå læk af modelparametre AI • Modificering af inputdata AI • Træn på manipuleret (*adversarial*) data AI • Begrænsning af inputdata AI • Øge validiteten af modeller m/hårde labels AI	• Penetrationstest af systemer • Monitorering af aktivitet • Overvågning af input og output AI • Benchmarking af model AI	• Nedlukning og isolation af inficerede systemer • Undersøgelse af angreb via logs • Backup-løsning

Kilde: BCG analyse; ekspertinterviews; CFCS; National Cyber Security Centre (NCSC); NIST; Center for Internet Security (CIS)

AI *AI-specifikt tiltag*