

Digitaliseringsstyrelsens vejledninger om de forbudte former for AI-praksis

Vejledning om forbud mod AI, der foretager social bedømmelse



Indhold

1. Indledende bemærkninger.....	2
2. AI-forordningens artikel 5, stk. 1, litra c	2
3. Betingelserne	2
4. Betingelse nr. 2: Om at evaluere/klassificere fysiske personer	3
5. Betingelse nr. 3: Om tidsperioden	4
6. Betingelse nr. 4: Om den sociale adfærd/personlighedstræk	4
7. Betingelse nr. 5: Om skadelig eller ugunstig behandling ...	5
8. Fiktivt eksempel på en forbudt AI-praksis efter AI-forordningens artikel 5, stk. 1, litra c.....	7

1. Indledende bemærkninger

For at sikre en bedre forståelse af indholdet i denne vejledning anbefales det, at den læses i sammenhæng med *"Hovedvejledning om de forbudte former for AI-praksis efter AI-forordningens artikel 5"*. Det skyldes, at hovedvejledningen indeholder en generel beskrivelse af de forbudte former for AI-praksis, mens denne vejledning udelukkende beskriver forbuddet mod AI, der foretager social bedømmelse (artikel 5, stk. 1, litra c). Hovedvejledningen kan findes på Digitaliseringsstyrelsens hjemmeside.

2. AI-forordningens artikel 5, stk. 1, litra c

Det følger af AI-forordningens¹ artikel 5, stk. 1, litra c, at det inden for EU's grænser er forbudt at bruge AI-systemer, som anvendes til social bedømmelse af fysiske personer, da det kan føre til forskelsbehandlende resultater og udelukkelse af visse grupper. Bestemmelsen har følgende ordlyd:

"Følgende form[er] for AI-praksis er forbudt:

omsætning, ibrugtagning eller anvendelse af AI-systemer til evaluering eller klassificering af fysiske personer eller grupper af personer over en given periode på grundlag af deres sociale adfærd eller kendte, udledte eller forudsagte personlige egenskaber eller personlighedstræk således, at den sociale bedømmelse fører til et af eller begge følgende udfald:

- i) skadelig eller ugunstig behandling af visse fysiske personer eller grupper af personer i sociale sammenhænge, som ikke har noget at gøre med de sammenhænge, i hvilke dataene oprindeligt blev genereret eller indsamlet;*
- ii) skadelig eller ugunstig behandling af visse fysiske personer eller grupper af personer, som er uberettiget eller uforholdsmæssig i forhold til deres sociale adfærd eller betydningen heraf"*

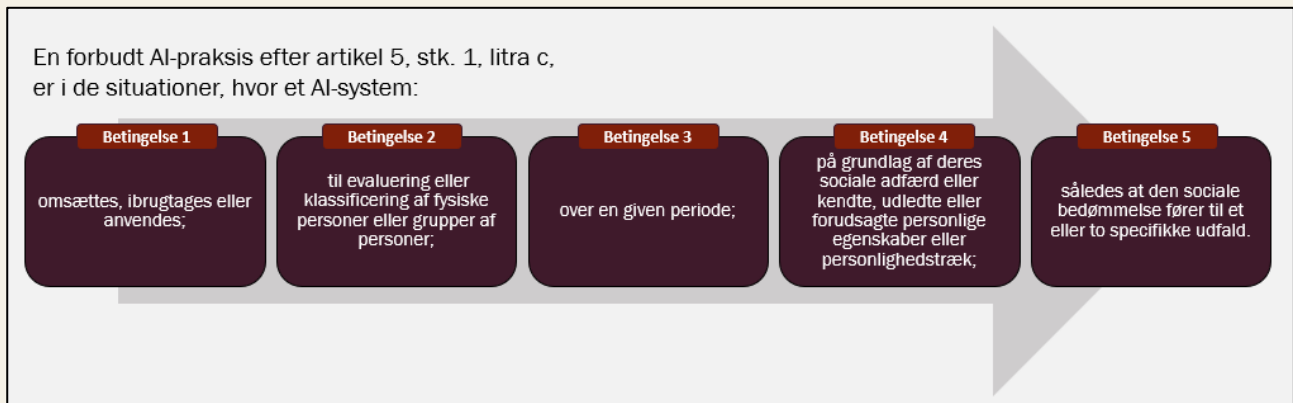
Formålet er at beskytte mennesker mod AI-systemer, der kan føre til forskelsbehandlende resultater og udelukkelse af visse grupper, hvilket kan krænke retten til værdighed og ikkeforskelsbehandling og værdierne lighed og retfærdighed².

3. Betingelserne

Af bestemmelsens ordlyd kan udledes, at der er fem betingelser, som *alle* skal være opfyldte, før et AI-system er forbudt. Det betyder også, at anvendelsen af et AI-system vil være lovligt, hvis én eller flere af betingelserne ikke er opfyldt. De fem betingelser fremgår af figur 1 nedenfor:

¹ Europa-Parlamentets og Rådets forordning (EU) 2024/1689 af 13. juni 2024 om harmoniserede regler for kunstig intelligens og om ændring af forordning (EF) nr. 300/2008, (EU) nr. 167/2013, (EU) nr. 168/2013, (EU) 2018/858, (EU) 2018/1139 og (EU) 2019/2144 samt direktiv 2014/90/EU, (EU) 2016/797 og (EU) 2020/1828 (forordningen om kunstig intelligens).

² AI-forordningens præambelbetragtning nr. 31.



Figur 1. De fem betingelser, som kan udledes af forbudskategorien i AI-forordningens artikel 5, stk. 1, litra c.

Da den første betingelse vedrørende "omsætning, ibrugtagning eller anvendelse af et AI-system" er beskrevet i afsnit 4.1 i hovedvejledningen, er det kun de fire øvrige betingelser, som vil blive uddybet i de følgende afsnit.

Det er vigtigt at slå fast, at størstedelen af de fem betingelser er meget nært forbundet, og det kan derfor også i praksis være svært at adskille dem fra hinanden.

4. Betingelse nr. 2: Om at evaluere/klassificere fysiske personer

Den anden betingelser som skal være opfyldt er, at AI-systemet skal anvendes til:

"evaluering eller klassificering af fysiske personer eller grupper af personer"

Betingelsen vil være opfyldt, hvis én af de nævnte anvendelsestyper bruges. Der er dermed ikke krav om, at AI-systemet anvendes til både evaluering og klassificering. I praksis kan det dog være svært at adskille evaluering og klassificering af fysiske personer, da de kan overlappe. Det er derfor ikke vigtigt, at man konkret er i stand til at fastslå, hvilken af de to anvendelsesformål, som AI-systemet faktisk anvender.

- **"Evaluering"** vil oftest omfatte en form for bedømmelse eller vurdering af en person eller grupper. Konceptet er beslægtet med "profilering", som følger af de databeskyttelsesretlige regler og udgør en specifik form for evaluering. Selvom der ikke refereres direkte til databeskyttelsesforordningen i forbudskategorien, kan det være et relevant sammenligningspunkt³.
- **"Klassificering"** henviser til det at gruppere eller kategorisere individer eller grupper baseret på specifikke karakteristika, såsom alder, køn eller højde, uden nødvendigvis at foretage en form for vurdering eller bedømmelse. I modsætning til evaluering, som indebærer en sådan vurdering, grupperer klassifikation blot

³ Ved "profilering" forstås efter databeskyttelsesforordningens artikel 4, nr. 4: "enhver form for automatisk behandling af personoplysninger, der består i at anvende personoplysninger til at evaluere bestemte personlige forhold vedrørende en fysisk person, navnlig for at analysere eller forudsige forhold vedrørende den fysiske persons arbejdsindsats, økonomiske situation, helbred, personlige præferencer, interesser, pålidelighed, adfærd, geografisk position eller bevægelser".

mennesker i henhold til definerede kriterier. Begrebet "klassifikation" dækker derfor bredere og omfatter kategoriseringer, der ikke kræver en analyse eller vurdering af en persons træk eller adfærd.

Forbuddet gælder både for offentlige og private aktører, men vedrører alene fysiske personer (individer eller grupper) og omfatter derfor ikke juridiske personer (virksomheder mv.).

Et eksempel på profilering fra databeskyttelsesretten findes i *SCHUFA I*-dommen, hvor et kreditvurderingssystem, der genererede en sandsynlighedsværdi (en score) for, om en person kunne tilbagebetale et lån, fandtes at udgøre profilering⁴. Fastlæggelsen af scoren var baseret på, at man ved at placere en person i en gruppe af andre personer med sammenlignelige karakteristika, som havde opført sig på en bestemt måde, kunne forudsige lignende adfærd.

5. Betingelse nr. 3: Om tidsperioden

Den tredje betingelse opstiller et krav om, at evalueringen eller klassificeringen skal ske:

"over en given periode"

Betingelsen indebærer, at en evaluering eller klassificering kun er omfattet af forbuddet, hvis den baserer sig på mere end blot et øjebliksbillede dannet ud fra data i en specifik kontekst. Vurderingen skal foretages ud fra et helhedsbillede, hvilket betyder, at den relevante periode kan variere afhængigt af den kontekst, hvori systemet anvendes. Der kan f.eks. være forskel på en social bedømmelse, der indgår i en adoptionssag, hvor vurderingen strækker sig over længere tid og en vurdering i forbindelse med en jobsamtale, hvor systemet evaluerer en kandidat på baggrund af et mere begrænset datagrundlag.⁵

6. Betingelse nr. 4: Om den sociale adfærd/personlighedstræk

Der stilles i den fjerde betingelse krav om, at AI-systemet evaluerer eller klassificerer personer:

"på grundlag af deres sociale adfærd eller kendte, udledte eller forudsagte personlige egenskaber eller personlighedstræk"

Betingelsen opstiller krav til de oplysninger, som AI-systemet anvender til brug for evalueringen eller klassificeringen af fysiske personer.

- Begrebet "**social adfærd**" dækker bredt og inkluderer handlinger, adfærd og vaner, som fysiske personer udviser i samfundet. Det kan f.eks. være deltagelse i kulturelle arrangementer, frivilligt arbejde, betalingshistorik, adfærd i forbindelse med brug af tjenester og samspil med offentlige eller private myndigheder, herunder overholdelse af love og regler. Ofte indsamles sådanne oplysninger fra mange forskellige datakilder og kombineres til én samlet vurdering.

⁴ Sag C-634/21.

⁵ EU-Kommissionens retningslinjer om forbudte former for praksis vedrørende kunstig intelligens i henhold til forordning (EU) 2024/1689 (AI-forordningen), afsnit 4.2.1. (b), s. 55-56 (den danske version).

- Begrebet **”personlige egenskaber eller personlighedstræk”** kan for eksempel omfatte køn, alder, seksuel orientering, etnicitet, indkomst, helbred, præferencer, interesser, indkomst, profession, bevægelsesmønstre, sundhedsdata, m.m. Der kan også være tale om personlighedsmæssige træk, som bruges til at danne profiler eller personlighedstyper.

Oplysningerne kan enten være indsamlet direkte fra personen selv eller igennem overvågning, tredjepartsdata eller udledt af andre informationer. De omfattede karakteristika kan være kendte, afledte eller forudsagte.

- Kendte oplysninger er typisk information, som personen selv har oplyst, f.eks. i en ansøgning eller brugerprofil.
- Afledte oplysninger er data som udledes fra andre datakilder, ofte af AI-systemet selv.
- Forudsagte oplysninger er AI-systemets skøn over personens karakteristika baseret på mønstergenkendelse.

7. Betingelse nr. 5: Om skadelig eller ugunstig behandling

For at et AI-system, som anvendes til social bedømmelse, er forbudt efter denne forbudskategori, skal det føre til en *skadelig* eller *ugunstig behandling* af personer. Betingelsen kan opfyldes, ved at AI-systemet har minimum ét af de to følgende udfald:

”i) skadelig eller ugunstig behandling af visse fysiske personer eller grupper af personer i sociale sammenhænge, som ikke har noget at gøre med de sammenhænge, i hvilke dataene oprindeligt blev genereret eller indsamlet

ii) skadelig eller ugunstig behandling af visse fysiske personer eller grupper af personer, som er uberettiget eller uforholdsmæssig i forhold til deres sociale adfærd eller betydningen heraf”

Ovenstående udfald kan godt være opfyldt samtidig. Betingelserne gennemgås enkeltvist nedenfor under ad) 1 og ad) 2).

Indledningsvist skal det dog bemærkes, at det er en forudsætning for både udfald 1 og 2, at der er en tydelig forbindelse (årsagssammenhæng) mellem den sociale bedømmelse, som AI-systemet foretager, og den måde, personen bliver behandlet på. Det betyder, at bedømmelsen skal være en direkte eller væsentlig årsag til den endelige vurdering, som bliver foretaget af AI-systemet. Dette gælder også, selvom AI-systemets bedømmelse kun delvist ligger til grund for den endelige beslutning, eller hvis der endnu ikke er sket skade, men AI-systemet er udviklet til eller i stand til at føre til skade.

Et AI-system, der anvendes i sammenspil med en menneskelig vurdering, er også omfattet, så længe den AI-genererede bedømmelse spiller en væsentlig rolle i den endelige beslutning. Forbuddet gælder desuden, selvom den organisation, der bruger bedømmelsen, ikke er den samme som den, der har produceret den.

Ad 1) Behandlingen sker i en anden social kontekst

Den første type udfald foreligger, når en person udsættes for skadelig eller ugunstig behandling baseret på en vurdering, der anvendes i en anden kontekst end den, hvor de underliggende oplysninger oprindeligt blev indsamlet i. Det kan f.eks. være, når irrelevante oplysninger om en persons internetvaner eller sociale aktiviteter bruges i en helt anden kontekst, såsom til skattekontrol eller vurdering af boligstøtte. Ofte vil dette ske uden personens viden eller forventning herom.

Fiktive eksempler på, at behandlingen sker i en anden social kontekst, kan f.eks. være:

- En skattemyndighed bruger oplysninger om borgeres sociale vaner (f.eks. at en borger ofte køber dyre designtasker eller ofte tager på ferie) til at udvælge dem til ekstra kontrol, selvom disse oplysninger ikke er relevante for den konkrete skattesag.
- En offentlig myndighed vurderer risikoen for socialt bedrageri ud fra oplysninger som etnicitet, internetbrug eller arbejdsvaner, selvom disse ikke har nogen tydelig sammenhæng med sagens formål.
- Et jobcenter anvender et AI-system, der blander relevante oplysninger om erhvervserfaring og uddannelse med irrelevante oplysninger om antal adresseskift eller køn for at vurdere, om en borger skal tilbydes støtte.

Ad 2) Behandlingen er ubegrundet eller ude af proportioner med den sociale adfærd

Den anden form for udfald foreligger, når en person udsættes for skadelig eller ugunstig behandling på baggrund af en social vurdering, hvor konsekvenserne ikke står i rimeligt forhold til den underliggende adfærd.

Her skal det vurderes, om der er balance mellem alvoren af personens handlinger og konsekvenserne af AI-systemets beslutning. Forbuddet kan også finde anvendelse, såfremt der slet ikke er noget sagligt eller legitimt grundlag for behandlingen.

Fiktive eksempler på uforholdsmæssig eller ubegrundet behandling kan bl.a. være:

- En offentlig myndighed registrerer familier som "risikofamilier" og fjerner børn fra hjemmet baseret på mindre overtrædelser, såsom at forældrene har modtaget en trafikbøde.
- En kommune bruger et AI-system til at sortere borgere i troværdighedskategorier og fratager dem adgang til offentlige ydelser, hvis de f.eks. ikke afleverer bibliotekets bøger til tiden eller har betalt en regning for sent.
- Et AI-system anvendes til at afsløre svindel med boligstøtte ved at inkludere irrelevante indikatorer, som internetforbrug, familierelationer eller uddannelsesniveau, hvilket fører til nægtelse af støtte uden sagligt grundlag.

Det er op til operatørerne af AI-systemet at dokumentere, at deres system kun anvender relevante og lovligt indsamlede oplysninger, og at den behandling, som de berørte personer bliver udsat for, er saglig, rimelig og forholdsmæssig. Tilsynsmyndighederne kan i den sammenhæng kræve, at operatørerne dokumenterer systemets funktion, hvilke data der anvendes, og hvorfor behandlingen er retfærdig.

8. Fiktivt eksempel på en forbudt AI-praksis efter AI-forordningens artikel 5, stk. 1, litra c

I dette afsnit gennemgås en konkret fiktiv case, hvori samtlige fem betingelser i bestemmelsen berøres. Beskrivelsen af casen er bygget op, så flere elementer vil blive tilføjet undervejs med henblik på at demonstrere, hvornår de enkelte betingelser individuelt er opfyldt, samt hvordan de hænger sammen indbyrdes.

Det er vigtigt at understrege, at denne case udelukkende vurderes efter forbudskategorien om skadelig social bedømmelse efter AI-forordningens artikel 5, stk. 1, litra c. Der bliver derfor ikke taget stilling til, hvorvidt hændelsesforløbet også vil være i strid med anden national eller EU-lovgivning, f.eks. strafferetten.

Case

Et større privat udlejningsfirma i Danmark udvikler og anvender et AI-system til intern risikovurdering af lejere og potentielle lejere i boligkomplekser. AI-systemet trækker over en længere periode på data fra offentligt tilgængelige kilder, sociale medier, adfærdsdata fra boligområderne (f.eks. hvem der kommer og går sent om aftenen), samt profilerede data fra datamæglere. AI-systemet beregner en "social adfærdsscore" på en skala fra 1-100, som opdateres løbende for hver person. Denne klassificering fører til:

- at potentielle lejere med lav score afvises automatisk.
- at nuværende beboere med lav score får begrænset deres adgang til fællesfaciliteter, og de overvåges tættere.

Systemets vurderinger baserer sig bl.a. på subjektive og indirekte indikatorer, fx "ofte ophold i områder med høj kriminalitet", "mange Facebook-opslag om fester", eller "tilknytning til personer med lav score".

Konsekvens: Flere personer – især personer fra bestemte bydele – bliver udelukket fra boliger, få begrænset deres adgang til fælles vaskeri i visse tidsrum og bliver blokeret fra brug af fælles festlokaler på baggrund af AI-systemets bedømmelse.

For at kunne vurdere om de forskellige betingelser i bestemmelsen er opfyldt, vil casen i det følgende blive brudt op. Casens forskellige delelementer fremgår af tekstbokse, mens forklaringer og uddybninger er skrevet med kursiv.

Vurdering af betingelse nr. 1) Om omsætning, ibrugtagning eller anvendelse af et AI-system:

Et AI-system bringes i omsætning i Danmark af større privat udlejningsfirma.

Det er først og fremmest en forudsætning for den videre vurdering, at AI-systemet er bragt i omsætning, ibrugtaget eller anvendt i EU. Det fremgår af casen, at AI-systemet er taget i brug i Danmark. Da anvendelse i et EU-land er tilstrækkeligt til, at AI-forordningen finder anvendelse, er betingelse nr. 1 opfyldt.

Det skal dernæst vurderes, hvorvidt AI-systemet anvendes til at evaluere eller klassificere mennesker.

Vurdering af betingelse nr. 2) Om at evaluere/klassificere fysiske personer:

Hvis der tilføjes: AI-systemet beregner på baggrund af forskellige data en "social adfærdsscore" på en skala fra 1-100, som opdateres løbende for hver person. Systemets vurderinger baserer sig bl.a. på subjektive og indirekte indikatorer, f.eks. "ofte ophold i områder med høj kriminalitet", "mange Facebook-opslag om fester" eller "personens tilknytning til personer med lav score".

I det AI-systemet indsamler data til brug for at beregne en "social adfærdsscore" på en skala fra 1-100, kan det udledes, at AI-systemet foretager en vis form for klassificering. Det forudsættes, at systemet tillægger de indsamlede data forskellige vægte og træffer en samlet vurdering og muligt en klassificering af personerne. Betingelse nr. 2 er derfor opfyldt.

Såfremt AI-systemet ikke tildelte personer en score fra 1-100 og dermed ikke foretog en vurdering af de indsamlede oplysninger, men i stedet blev anvendt til indsamling af oplysninger til brug for en senere menneskelig vurdering, så ville AI-systemet ikke længere foretage en vurdering eller klassificering af personer og vil derfor ikke opfylde betingelse nr. 2. Anvendelsen af systemet ville dermed ikke være forbudt efter AI-forordningen.

Dernæst skal det vurderes, om evalueringen eller klassificeringen sker over en given periode.

Vurdering af betingelse nr. 3) Om tidsperioden:

Hvis der tilføjes: AI-systemet trækker løbende på data fra offentligt tilgængelige kilder.

I casen angives det, at AI-systemet baserer sin vurdering på data fra offentligt tilgængelige kilder over en længere tidsperiode. Det betyder også, at AI-systemet ikke træffer sin vurdering alene baseret på et øjebliksbillede og betingelsen er derfor opfyldt.

Såfremt AI-systemet kun vurderede nye ansøgere én gang, når de pågældende ansøgte om bolig, og at AI-systemet alene tog afsæt i oplysninger fra, som ikke efterfølgende ville blive opdateret eller overvåget, ville betingelsen om evaluering eller klassificering over en given periode i dette tilfælde ikke være opfyldt, da der er tale om et enkeltstående tilfælde. Anvendelsen af AI-systemet ville dermed ikke være forbudt efter AI-forordningen.

Dernæst skal det vurderes, om evalueringen eller klassificeringen sker på grundlag af personernes sociale adfærd eller kendte, udledte eller forudsagte personlige egenskaber eller personlighedstræk.

Vurdering af betingelse nr. 4) Om den sociale adfærd/personlighedstræk:

Hvis der tilføjes: AI-systemet trækker særligt data fra sociale medier, adfærdsdata fra boligområderne (f.eks. hvem der kommer og går sent om aftenen) samt profilerede data fra datamæglere.

Det fremgår af casen, at AI-systemet benytter sig af en række forskellige datakilder, der er udtryk for personens sociale adfærd, til at foretage vurderingen. Digitale fodspor på sociale medier og adfærdsdata vil i de fleste tilfælde opfylde betingelsen om, at evalueringen eller klassificeringen sker på grundlag af personernes sociale adfærd eller

kendte, udledte eller forudsagte personlige egenskaber eller personlighedstræk. Casen indikerer hertil, at der både anvendes kendte og udledte personlige egenskaber. Betingelse nr. 4 er derfor opfyldt.

Det skal understreges, at begreberne "social adfærd", "personlige egenskaber" og "personlighedstræk" dækker meget bredt, hvorfor det generelt må antages, at de fleste AI-systemer, som foretager vurderinger af fysiske personer, vil opfylde denne betingelse.

Endelig skal det vurderes, om den sociale bedømmelse fører til skadelig eller ugunstig behandling af visse fysiske personer.

Vurdering af betingelse nr. 5) Om skadelig eller ugunstig behandling:

Hvis der tilføjes: AI-systemets tildeling af en "social adfærdsscore" på en skala fra 1-100 på baggrund af de forskellige datakilder fører bl.a. til: At potentielle lejere med lav score afvises automatisk, at nuværende beboere med lav score får begrænset deres adgang til fællesfaciliteter samt, at nuværende beboere overvåges tættere. Det har den konsekvens, at flere personer – især personer fra bestemte bydele – bliver udelukket fra boliger, få begrænset deres adgang til fælles vaskeri i visse tidsrum og bliver blokeret fra brug af fælles festlokaler på baggrund af AI-systemets bedømmelse.

De forskellige elementer i casen indebærer, at konsekvenserne af den sociale bedømmelse kan være omfattet af begge de to mulige udfaldsmuligheder efter bestemmelsen, idet:

- i) Behandlingen sker i en anden sammenhæng end den, hvor dataene oprindeligt er indsamlet. Da dataene bruges til at træffe beslutning om boligudlejning og adgang til fælles tilbud, og dermed i en helt anden kontekst end den, hvor dataene oprindeligt blev genereret, er betingelsen i bestemmelsens artikel 5, stk. 1, litra c, i) opfyldt.
- ii) Behandlingen er uberettiget eller uforholdsmæssig i forhold til den sociale adfærd. Casen beskriver, at personer fra visse postnumre bliver udelukket fra boliger, hvilket vil være et uforholdsmæssigt og uberettiget udfald. Casen opfylder derfor også betingelsen i bestemmelsens artikel 5, stk. 1, litra c, ii).

Det skal understreges, at betingelsen også ville være opfyldt, såfremt udfaldet alene faldt ind under enten i) eller ii). Der er ikke krav om, at begge skal være opfyldt.

Såfremt AI-systemet vurderede den enkelte lejers adgang til at låne bygningens festlokaler baseret på lejerens historik, herunder anmeldelser fra rengøringsfolk, støjklager og skade på inventar - det vil sige, at der på baggrund af disse datakilder foretages en evaluering af, hvor stor sandsynligheden er for, at lejerer vil efterlade lokalet i en uacceptabel tilstand – ville betingelsen ikke være opfyldt.

I denne situation vil AI-systemet analysere kendte data, som vil have en nær tilknytning til og sammenhæng med det endelige udfald. Her kan konsekvensen ikke anses for værende uproportionel. Den sidste betingelse om skadelige eller ugunstig behandling ville i dette tilfælde ikke være opfyldt, og anvendelsen af AI-systemet ville derfor ikke være forbudt efter AI-forordningen, selvom de øvrige betingelser (nr. 1-4) er opfyldt.

