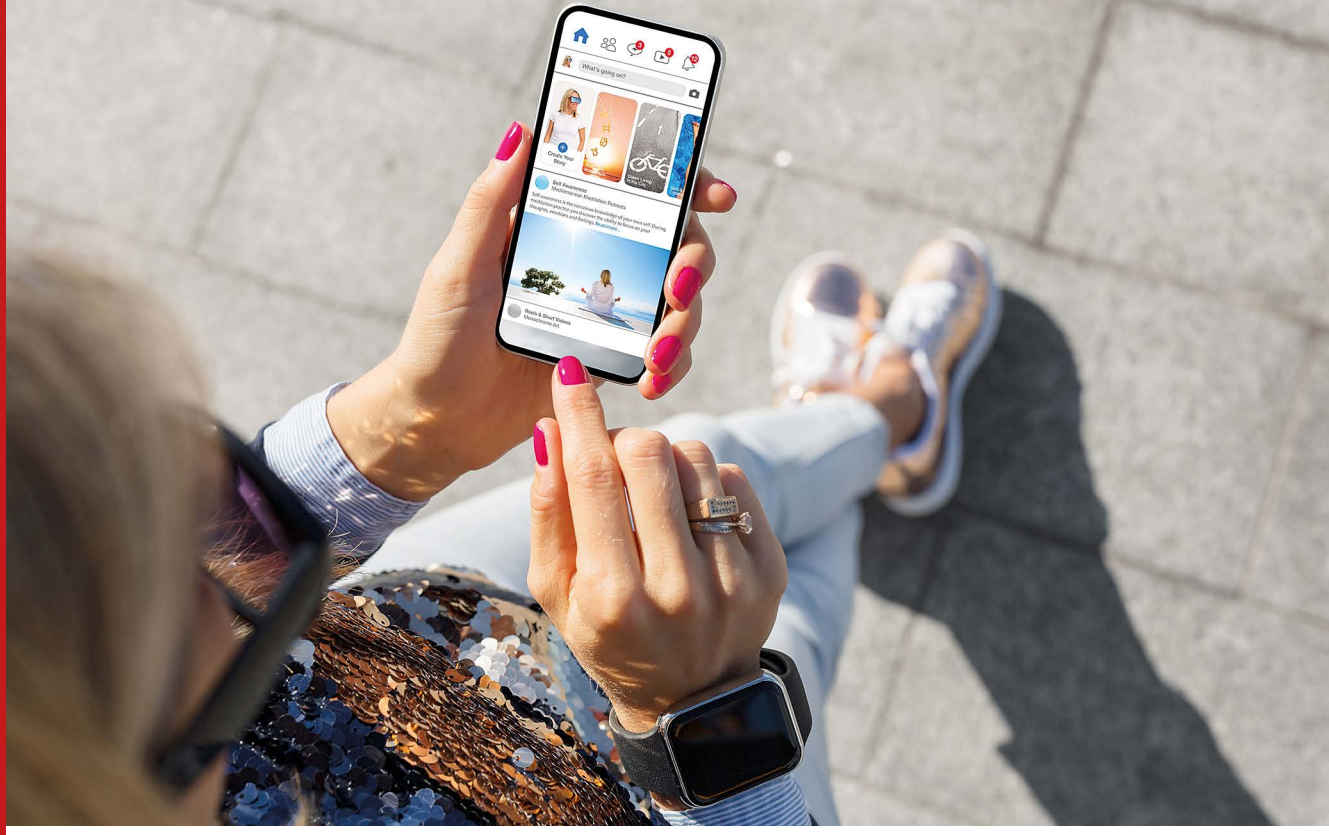
De forklaringer, du får, når du trykker "hvorfor ser jeg denne annonce?" på Facebook, baserer sig på de oplysninger, du har delt med platformen samt dine interaktioner på Facebook og eventuelt også andre sider. Et studie fra 2018 viste dog, at forklaringerne typisk er mangelfulde og vage, idet de i mange tilfælde undlader de dele af dataindsamlingen, der kan føles intimiderende for brugerne. Det giver anledning til kritiske spørgsmål om, hvad forklaringernes formål reelt er (Andreou et al. 2018).
Foto: Shutterstock



KUNSTIG INTELLIGENS

og kunsten at give en god forklaring

Hvad er en god forklaring? Selv om svaret på dette spørgsmål ikke er entydigt, rummer det flere vigtige betragtninger for udviklingen af såkaldt Explainable AI, der er metoder til at give forståelige oversættelser mellem kunstig intelligens og mennesker.

Forfatter:
Torben Esbo Agergaard er ph.d.-studerende ved Center for Videnskabsstudier, Institut for Matematik, Aarhus Universitet. Torbens forskning undersøger etiske og erkendelsesmæssige udfordringer ved Explainable AI, og hvordan vi i fremtiden kan få bedre oversættelser mellem systemer baseret på kunstig intelligens og mennesker.
ta@css.au.dk

Kunstig intelligens (eller blot AI efter det engelske Artificial Intelligence) oplever som bekendt en rivende udvikling og udbredelse i disse år. Det skyldes i høj grad udviklingen af meget komplicerede modeller, der med mindre og mindre hjælp fra mennesker kan udlede mønstre i store mængder data. Men succesen for kunstig intelligens er forbundet med et voksende problem: Når modellerne så at sige lærer fra data på egen hånd, bliver det sværere for brugerne, ja, selv for udviklerne, at forstå, hvad modellerne lægger vægt på, når de lærer. Modellerne er blevet uigennemskuelige; de er blevet til såkaldte "black boxes", og det gør det eksempelvis svært at korrigere fejl i dem.

Det er netop med ønsket om at imødekomme udfordringen med "black box"-modeller, at forskere og IT-udviklere lige nu arbejder på højtryk for at udvikle metoder, der gør det muligt at forklare, hvordan modellerne når frem til deres output. Forklaringerne kan tage form som tekst, grafik eller statistik og kan eksempelvis fremhæve, hvilke variabler modellerne har lagt vægt på i en given situation. Det kunne være hvilke pixels, der har været udslagsgivende ved vurderingen af et scanningsbillede, eller hvilke interaktioner Instagrams algoritme lægger vægt på, når du får serveret reklamer for hvide sneakers i dit feed.

Metoder, der forklarer, hvorfor modellerne når frem til deres output, kalder vi Explainable Artificial Intelligence (oftest forkortet XAI). Kort sagt er XAI metoder til at oversætte mellem AI-baserede systemer og mennesker.

Bliver din ansøgning om et banklån afvist, kan XAI hjælpe med at forklare hvorfor, hvis beslutningen er baseret på AI, hvilket er udbredt i dag. På Facebook og Instagram kan du bede om forklaringer på, hvorfor deres marketingsalgoritme har puttet en given annonce i dit feed. Og når et AI-baseret GPS-system som Google Maps fortæller, at en udvalgt rute er "den hurtigste", er det også en form for XAI, fordi det er en forklaring på, hvorfor den givne rute er valgt frem for andre.

Et forklaringsproblem

For at XAI-metoderne kan få succes, kræver det svar på et spørgsmål af mere filosofisk end teknisk karakter: Hvad er en god forklaring?

Lige så simpelt som spørgsmålet lyder, lige så kompliceret er svaret. Der findes nemlig ingen almene kriterier for, hvad der tæller som en god forklaring. I stedet afhænger det af den sammenhæng, som forklaringen indgår i, herunder hvilke mål modtageren skal opnå.

Det er ellers fristende at prøve at opstille almene kriterier for den gode forklaring. Det kunne være kriterier som korrekthed og nøjagtighed. Korrekthed er dog ikke et tilfredsstillende kriterium. Der findes nemlig masser af korrekte forklaringer på et givet fænomen, men de er ikke lige informative. Går der ild i dit lokale bageri, er forklaringen "der var ilt i luften til at nære ilden" lige så korrekt som "bageren var begyndt at eksperimentere med at bruge sprit i stedet for olie i dejen". Den er bare ikke lige så interessant, vel?

Og hvad nøjagtighed angår, er det faktisk sjældent, at en fuldstændig nøjagtig forklaring er mulig eller

Centrale begreber indenfor kunstig intelligens

Kunstig intelligens

En gængs opfattelse er, at der er tale om kunstig intelligens, når et computerbaseret system løser opgaver via en fremgangsmåde, der minder om menneskelig intelligens. Det kunne være at opstille analyser, vurderinger eller forudsigelser på baggrund af data, for eksempel at analysere brugeradfærden på et socialt medie eller streamingtjeneste med henblik på at vise annoncer eller anbefalinger, eller at vurdere symptombilledet hos en patient. I andre situationer kan kunstig intelligens omsætte data til handlinger eller interaktioner med mennesker, for eksempel chatbots.

Algoritmer

En algoritme er en sekvens af trin, der beskriver, hvordan man på baggrund af et input når frem til et resultat, der løser en opgave. En bageopskrift er en slags algoritme, hvor input er sukker, mel, mælk m.v. og resultatet er en kage. I en computermæssig sammenhæng beskriver algoritmer typisk, hvordan systemet skal omsætte et data-input til et resultat. Selv om algoritmer er centrale for kunstig intelligens, er relationen ikke en-til-en. Et system baseret på kunstig intelligens anvender ofte flere algoritmer.

Model

En model er en matematisk (for eksempel formelbaseret eller grafisk) beskrivelse af, hvordan vi kan strukturere data, og hvad vi kan forvente af fremtidige data, hvis beskrivelsen er god. En lineær eller eksponentiel fremskrivning er eksempler på simple modeller. Modeller er dog sjældent perfekte og vil derfor ofte være forbundet med en vis fejlmargin eller usikkerhed.

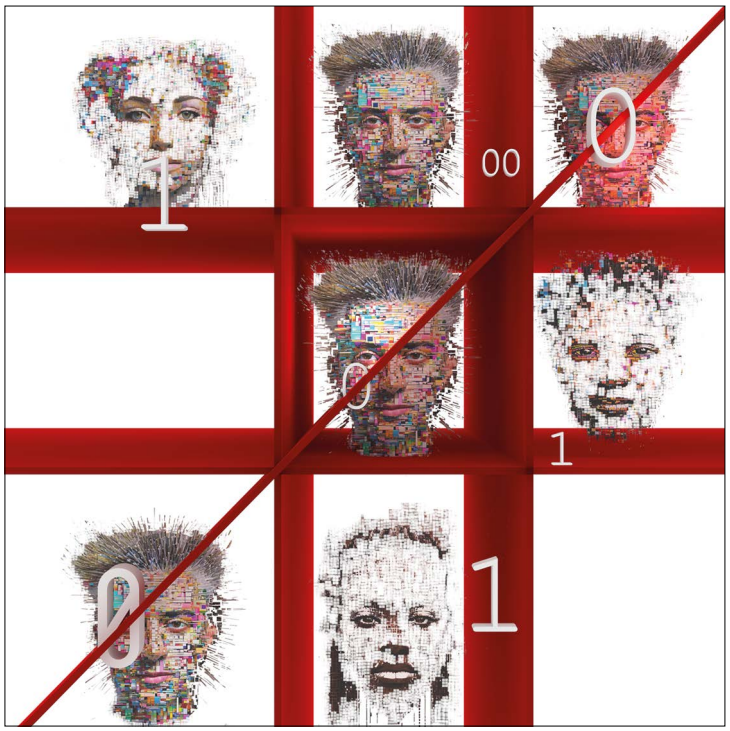
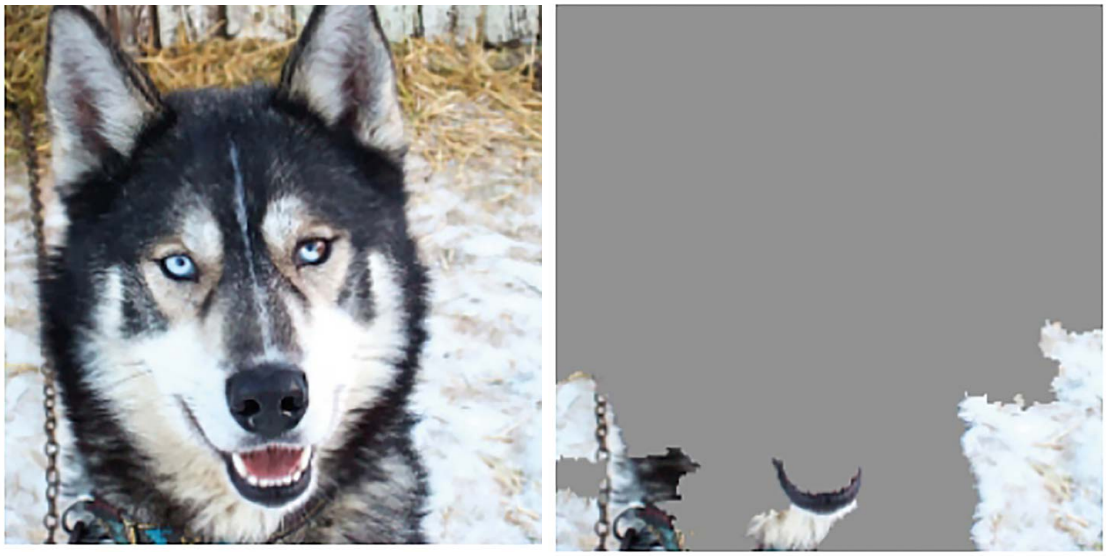


Illustration: Amritha R Warrier & AI4Media via betterimageofai.org

I 2016 viste en undersøgelse af en algoritme brugt i det amerikanske retssystem til at vurdere risikoen for, at en tiltalt begik kriminalitet igen, at tiltalte med mørk hudfarve systematisk fik højere risiko-score end personer med lys hudfarve tiltalt for samme type forbrydelse. Det er et eksempel på endnu en type af problemer med uigennemskuelige AI-systemer, nemlig såkaldte biases, hvor anvendelsen af systemet fører til diskrimination af bestemte befolkningsgrupper.



Modellen her skulle lære at skelne mellem en husky og en ulv, men fordi ulvebillederne i træningssættet havde sne i baggrunden, forbandt modellen ulve med sne, og huskyen på billedet til venstre blev derfor klassificeret som ulv. XAI-forklaringen på billedet til højre blotlægger fejlen ved at fremhæve pixels i sneen frem for karakteristika ved den fejlklassificerede husky. Kilde: Ribeiro et al. (2016, s. 1143).

ønskværdig. Tænk, hvis din GPS forklarede alle skridt i beregningerne bag dens rutevejledning. Så hellere gå tilbage til at finde vej via et landkort.

Det hele bliver mere kompliceret af, at den gode forklaring heller ikke afhænger af, hvilket spørgsmål vi stiller, men i hvilken sammenhæng vi stiller dem i. Hvad der er et godt svar på spørgsmålet "hvordan virker mit TV?", afhænger eksempelvis af, om du ønsker at tænde det, eller om du forsøger at reparere det, hvis det ikke fungerer.

Vi er derfor tilbage ved, at en forklaring er god, hvis den gør det muligt at opnå de mål, der er vigtige i den givne sammenhæng, for eksempel at du kan reparere dit TV.

Dette efterlader os imidlertid med to nye spørgsmål: 1) Hvordan konstruerer og udvælger vi forklaringer, der gør modtageren i stand til at opnå disse mål? Og 2) Hvad er egentlig målene i en given sammenhæng?

Begge spørgsmål er naturligvis også vigtige for udviklingen af XAI og giver anledning til en række etiske nøgleudfordringer for feltets udvikling. Dem skal vi se nærmere på nu.

At designe forklaringer er en balancegang

Oversætter vi det første spørgsmål til en XAI-sammenhæng, lyder det noget i stil med: Hvordan kan udviklere designe forklaringer, der gør det muligt for brugeren at opnå de mål, der er vigtige i en given anvendelsessammenhæng? Svaret er, at udviklerne skal finde en balance mellem brugerens informationsbehov og de ressourcer, hun har til rådighed i situationen. Ressourcer kan være, hvor meget tid, koncentration eller forudgående viden brugeren har.

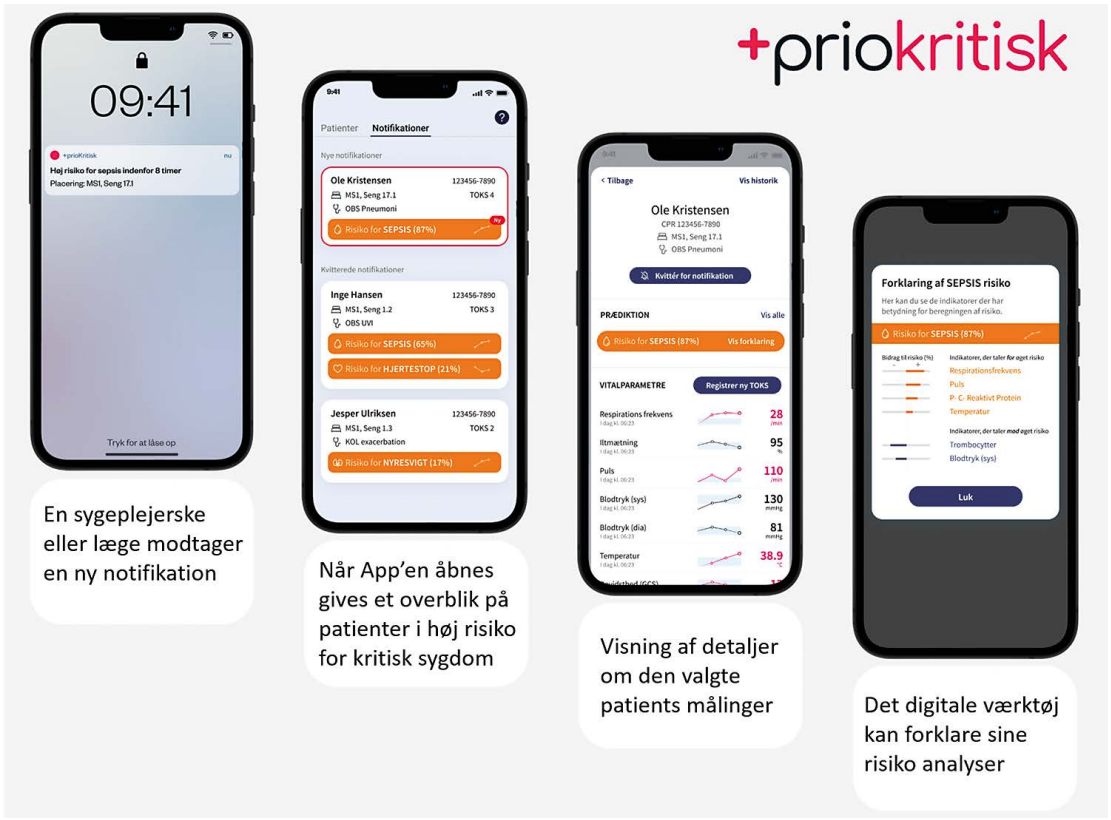
Tag GPS'en i din bil, der har forklaret valget af rute med, at det er "den hurtigste". Den kunne sagtens have givet en længere forklaring om, hvad den har antaget om din kørerstil og ventetiden i lyskrydsene på ruten. Det kunne måske gøre dig i stand til at korrigere ruten til en endnu hurtigere, hvis du ved, at du kører anderledes, end GPS'en antager, eller at der er lyskryds på ruten, der skifter umanerlig langsomt. Men når du sidder i trafikken, er der ikke tid til lange forklaringer. Hellere komme to minutter senere frem end at risikere at køre galt, fordi du bliver overdynget med information.

I andre tilfælde er det dog sværere at vægte informationsbehov og ressourcer, fordi der er tunge risici på begge sider. Indenfor lægevidenskaben vinder AI-baserede systemer frem, og det samme gør også XAI-metoder til at understøtte, at lægerne får forklaret systemernes vurderinger. Her er det vigtigt, at lægerne får deres informationsbehov dækket for at kunne være kritiske over for systemets vurderinger for patienternes helbreds skyld. Men også læger har travlt, og komplekse eller utidige forklaringer kan medføre spild af værdifuld tid.

At designe forklaringer, der balancerer informationsbehov og ressourcer, er en nøgleudfordring for XAI. Og netop fordi AI kommer krybende ind i flere og flere situationer i vores liv, er det vigtigt, at der også fremover forbliver stor fokus på at forstå de behov og ressourcer, situationerne fordrer. Men "den gode forklaring" er ikke kun et designproblem. På et mere fundamentalt niveau er vi nødt til at sikre, at selve formålet med forklaringerne faktisk er i brugernes interesse. Vi vender os derfor nu mod det ofte oversete spørgsmål 2) fra tidligere:

Hvad er egentlig målene i en given sammenhæng?

At GPS'en skal få os hurtigt men stadig sikkert frem, og at læger skal træffe de bedste beslutninger for patienternes helbred, om end



Eksempel på anvendelse af XAI i sundhedsvæsnet. SOFUS (tidl. +priokritisk) giver ikke blot en risikoscore for at indlagte patienter får akutte sygdomme som blodforgiftning (sepsis). Den giver også en forklaring på risikovurderingen ved at fremhæve de symptomer, der er mest udslagsgivende for vurderingen (billede 4) SOFUS er under implementering på Regionshospitalet Horsens.
Kilde: Regionshospitalet Horsens hjemmeside.

tidspres også er en faktor, virker som oplagte og rimelige mål for de XAI-baserede forklaringer, vi har kigget på. I mange situationer virker målene med forklaringerne da også så oplagte og rimelige, at vi ikke behøver at sige dem højt. Kradser vi i overfladen, vil vi i nogle situationer dog finde, at afsenderen kan have skjulte mål med forklaringerne, der ikke er oplagte for modtageren, hvilket kan medføre etiske problemer. I en XAI-sammenhæng kan afsenderen både være udvikleren eller et firma eller en organisation, der tager en algoritme i brug.

Tag eksemplet, hvor du har fået din ansøgning om et banklån afvist med forklaringen "du tjener ikke penge nok lige nu". Det virker umiddelbart som god stil at give en forklaring som denne på algoritmens vurdering, idet det giver dig mulighed for at reagere ved for eksempel at gå op i tid på dit arbejde.

Men forklaringen kunne lige så godt være "du bruger for mange penge". Et nedsat forbrug vil kunne give dig lige så mange penge på din bankkonto som det at tjene flere penge og burde derfor også kvalificere dig til lånet. Og har du svært ved at gå op i tid på dit arbejde, er denne forklaring formentligt bedre for dig end den forrige. Men hvis din bank tjener penge på gebyrer, hver gang du bruger penge, er det en bedre forretning for dem, at du tjener og bruger flere penge, end at du sparer dem op. Dermed kan banken have en interesse i, at du får den første frem for den anden forklaring.

På et mere overordnet plan kan forklaringen også være med til at flytte fokus hen på, hvad du skal gøre for at gøre dig fortjent til lånet. Her er det dog værd at huske, at algoritmens vurdering baserer sig på de regler for udlån, som i sidste ende er bankens ansvar. Men for banken kan det være mere bekvemt at lægge fokus på kundens ansvar frem for deres eget.

Eksemplet viser, hvordan forklaringer på subtile måder kan (mis) bruges til andre formål end at give modtageren forståelse. Selv korrekte forklaringer kan som i bankeksemplet være udvalgt, så de nudger modtageren til den adfærd, som afsenderen ønsker, eller være afgrænset til kun at forklare de årsager, der er i afsenderens interesse.

Ligesom AI-modeller ikke er et stykke neutral matematik, men baserer sig på aktive beslutninger om, hvad de skal og ikke skal kunne, er det samme gældende for de forklaringer, der oversætter mellem algoritmerne og deres brugere.

Den anden nøgleudfordring for XAI er derfor at sikre, at forklaringerne faktisk er til gavn for brugerne og ikke misbruges til formål, der kun er i afsenderens interesse. Dette kræver, at der fremover bliver lagt mere vægt på at forstå og blotlægge de fordækte mål, som forklaringerne kan misbruges til at understøtte. ■

Om artiklen:
Artiklen udløber af projektet TREAT (Towards Responsible Explainable AI Technologies), der er ledet af lektor Rune Nyrup og støttet af Danmarks Frie Forskningsfond. Der skal lyde en stor tak til Rune og til Kristian Hvidtfelt Nielsen, centerleder for Center for Videnskabsstudier, for værdifuld sparring i forbindelse med artiklen.

Videre læsning:
Du finder en god og frit tilgængelig introduktion til Explainable AI i bogen her:
Molnar, C. (2024). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Interpretable Machine Learning.

Husky-ulve-billedet er fra denne indflydelsesrige artikel om LIME, en af de mest udbredte XAI-metoder:
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. doi.org/10.1145/2939672.2939778.

Læs mere om SOFUS og følg med i implementeringen her:
Forskningsenheden Regionshospitalet Horsens. (n.d.). SOFUS: Sequential Organ Failure sUpport System. www.sofus-ai.com/home-dk.

Omtalte studie af Facebooks "hvorfor ser jeg denne annonce?"-forklaringer:
Andreou, A., Venkatadri, G., Goga, O., Gummadi, K. P., Loiseau, P., & Mislove, A. (2018). Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations. Proceedings of the Network and Distributed System Security Symposium (NDSS 2018). 1–15. doi.org/10.14722/ndss.2018.23204.