48-timers hjemmeprøve i Datalogiens Videnskabsteori (VtDat)

Ordinær eksamen 2024

Eksamensnummer: 102

Datalogisk Institut
Københavns Universitet
Danmark
June 14, 2024

Contents

Spørgsmål 1

Stephen Hilgartner snakker om tre typer ansvar i forbindelse med videnskabelig uredelighed: kausalt ansvar, moralsk ansvar og politisk ansvar. Kausalt ansvar handler om årsagerne til problemet. Hilgartner siger, at det ikke kun er den enkelte forskers skyld, men også større systemiske problemer i det videnskabelige miljø. For eksempel er der pres for at udgive mange artikler og mangel på ordentlig opsyn af studerende [1].

Moralsk ansvar handler om, hvem der er skyld i uredelighed. Ifølge Hilgartner bør ikke kun den enkelte forsker holdes ansvarlig, men også medforfattere og vejledere, der ikke gør deres arbejde ordentligt. Universiteter og forskningsinstitutter, der ikke håndterer anklager om uredelighed effektivt, kan også være medskyldige. Videnskabelig uredelighed sjældent er resultat af en enkelt persons handlinger, men ofte et systemisk problem [1].

Politisk ansvar handler om, hvem der skal tage handling mod uredelighed. Hilgartner siger, at der traditionelt set ikke har været så meget fokus på politisk ansvar, men at det nu er nødvendigt med mere omfattende løsninger [1].

Hilgartner foreslår fire måder at tackle problemet på:

- Håndhævelse af loven: Her skal der være fokus på at opdage og straffe uredelighed hårdt, så forskere bliver afskrækket fra at snyde. Dette indebærer effektiv undersøgelse af anklager, hurtig og hård straf til de skyldige samt beskyttelse af whistleblowers mod gengældelse. I Danmark er denne tilgang implementeret gennem Nævnet for Videnskabelig Uredelighed (NVU), der har til ansvaret for at håndtere sager om videnskabelig uredelighed som fabrikering, forfalskning og plagering [1, 2].
- Tilsyn: Det handler om at øge kontrol med forskerne ved at stille krav til, hvordan data skal opbevares. Foreslagene inkluderer krav om bedre registrering og opbevaring af data og regelmæssige dataaudits. Højere krav om publicering og opbevaring af rådata og forsøgsdesign er nu almindelige praksis, hvilket er en direkte anvendelse af Hilgartners forslag [1, 2].
- Uddannelse: Forskere skal uddannes i god videnskabelig praksis, så de ikke bliver fristet til at snyde og i stedet opbygger en kultur af integritet og god etik. Foreslagene inkluderer mere intensiv interaktion mellem senior forskere og studerende og grundigere undervisning i forskningsmetoder og etik. Dette er blevet implementeret gennem kurser i "Fagets Videnskabsteori", som styrker uddannelsen i god videnskabelig praksis [1, 2].
- Belønningssystem: Der skal sættes mere fokus på kvalitet over kvantitet, så der er mindre incitament til at snyde. For eksempel foreslog Harvard Medical School, at kvaliteten af publikationer vægtes mere. Det er usikkert, om belønningssystemerne er ændret nok til at fremme god videnskabelig praksis. Dette spørgsmål kan kobles til systemisk etik, hvor strukturer påvirker forskernes moralske valg [1, 2].

Samlet set viser Hilgartners teori, hvordan systemisk forandring og individuel ansvarlighed kan kombineres for at fremme etisk forskning og minimere uredelighed. Ved at arbejde med disse principper kan forskningsmiljøet blive mere integritetsbaseret og ansvarligt. Det kræver en koordineret indsats fra både enkeltpersoner og institutioner for at skabe en kultur, hvor ærlighed og etik er i centrum, og hvor strukturerne understøtter god videnskabelig praksis.

Spørgsmål 2

Forklarbarhed er essentielt for AI-modeller, da det handler om at kunne forstå og forklare, hvordan og hvorfor en AI-model træffer sine beslutninger. Data indeholder ofte mange variabler, og det kan være svært at skelne mellem de variabler, der faktisk betyder noget for modellens forudsigelser, og dem, der blot er støj. Dette ses især med de såkaldte "black box"-modeller, hvor det kan være næsten umuligt at forstå, hvorfor modellen træffer bestemte beslutninger [3]. Det er ikke kun vigtigt at være teknisk præcis, det er også nødvendigt at skabe forklaringer, som mennesker kan forstå og acceptere. En undersøgelse foretaget af AI-forskere viser, at for at gøre AI-modeller mere forståelige, er det nødvendigt at udvikle AI-modeller, der er baseret på indsigter fra samfunds- og kognitionsvidenskab [4].

"Korrelation medfører ikke kausalitet" er en klassisk udfordring i statistik og dataanalyse. Det er en fejl at antage, at korrelation mellem to ting indebærer en direkte årsagssammenhæng. Dette problem er særligt relevant i ML-modeller, der ofte identificerer mønstre og korrelationer i data uden at kunne skelne mellem reelle årsagsforhold. Dette problem ses med "spurious correlations", hvor der kan opstå tilfældige sammenhænge uden nogen reel kausal relation. Hvis disse modeller anvendes uden kritisk at evaluere de identificerede mønstre, kan det føre til fejlagtige konklusioner [3, 2].

Algoritmisk bias opstår, når AI-systemer viser bias i data, design eller anvendelse. Når ML-modeller bruges til beslutningsprocesser eller rådgivning, er det vigtigt at huske på, at forudsigelser ofte baseres på historiske data. Dette kan føre til et induktionsproblem, hvor modellerne kun gentager de mønstre, de er trænet på, og dermed viderefører tidligere biases ind i fremtiden. Derfor er det vigtigt at være opmærksom på, hvor data kommer fra, og hvordan de er blevet udvalgt, for at minimere risikoen for algoritmisk bias [3].

Spørgsmål 3

Når man taler om de epistemiske udfordringer ved brugen af AI og maskinlæring, er det vigtigt at forstå, hvordan disse udfordringer kan føre til etiske problemer. Dette afsnit vil udforske nogle af disse problemer baseret på kursuslitteraturen.

Automation bias beskriver tendensen til at have for meget tillid til automatiserede systemer. Dette er særligt problematisk i militære miljøer, hvor AI-drevne beslutninger ofte anses for mere pålidelige end menneskelige vurderinger. Det kan føre til situationer, hvor soldater blindt følger AI-systemernes anbefalinger, selvom de er fejlagtige eller præget af bias. For eksempel kan AI i våbensystemer fejlagtigt identificere civile som fjender. Det er vigtigt at kunne forklare beslutninger præcist og sikre, at beslutningerne er nøjagtige og pålidelige, hvilket er et centralt kriterium for at vurdere etikken omkring autonome våben [5]. Manglende forklarbarhed i AI-systemers beslutningsprocesser udgør derfor en væsentlig etisk udfordring, især når det kommer til LAWs. Uden en klar forståelse af AI's beslutningsmekanismer kan operatørerne hverken verificere eller udfordre systemets handlinger, hvilket gør det svært at holde nogen ansvarlige for fejl. Bekymringen er, at AI-våbensystemer kan handle på måder, der overskrider menneskelige forventninger, hvilket komplicerer spørgsmål om ansvar og kontrol [5, 6]. Man kan argumentere for, at den sidste beslutning altid bør træffes af en person for at sikre en grundig menneskelig bedømmelse og reducere risikoen for katastrofale fejl. Desuden kan Explainable AI (XAI) inddrages for at hjælpe operatørerne ved at give dem en grundig forklaring på AI-modellernes forslag, hvilket vil styrke beslutningsgrundlaget og øge ansvarligheden [4].

Algoritmisk bias opstår, når bias indføres i data, design eller anvendelse af AI. Mange ML-modeller trænes på historiske data, som kan videreføre og forstærke tidligere bias. Brugen af AI-modeller til at oprette lister over potentielle terrorister er problematisk. Det kan forstærke eksisterende bias og føre til fejl, hvilket kan resulterer i angreb på uskyldige mennesker. Der er derfor stor forskel på hvilken sammenhæng AI bliver brugt i, når det kommer til etikken. Man kan overveje kun at bruge LAWs i forsvarssituationer og ikke til angreb. I forsvarssituationer vil der være færre etiske dilemmaer, da det er nemmere at identificere fare fra missiler end fra mennesker [6].
Et eksempel på algoritmisk bias inden for det juridiske system er COMPAS-algoritmen i det amerikanske retssystem, der anvendes til at vurdere risikoen for tilbagefald blandt kriminelle. Undersøgelser har vist, at algoritmen systematisk favoriserer hvide personer og fejlagtigt vurderer farvede personer som højrisiko, selvom de ikke begår ny kriminalitet. Dette skaber en konflikt mellem effektivitet og retfærdighed, da algoritmen forstærker eksisterende uligheder og diskrimination. Dette understreger behovet for at evaluere og revidere algoritmer for at sikre, at de ikke viderefører eller forstærker systemiske bias [2].

Et klassisk eksempel på problemet med "korrelation medfører ikke kausalitet" ses i Google Flu Trends-projektet. Google Flu Trends blev lanceret for at bruge søgedata til at forudsige influenzaspredning ved at analysere millioner af søgeforespørgsler og finde mønstre, der korrelerede med influenzasæsoner. Projektet begyndte at fejle, fordi modellen antog, at korrelationerne mellem søgeforespørgsler og influenzatilfælde var kausale. Faktorer som øget mediedækning af influenzasæsonen fik flere mennesker til at søge efter influenzarelaterede symptomer, selvom de ikke var syge. Induktionsproblemet i Google Flu Trends handlede om udfordringen ved at forudsige fremtidige hændelser baseret på historiske data. Modellen, der blev trænet på data fra 2003-2008, kunne ikke forudsige en usædvanligt tidlig influenzasæson i 2009, fordi den var afhængig af mønstre i træningsdataene. Denne fejltolkning førte til overestimering af influenzatilfælde og kunne potentielt have skadet folkesundheden ved at skabe falsk alarm eller overse reelle udbrud [7].

Spørgsmål 4

Kausalt ansvar i forbindelse med udviklingen af LAWS indebærer at identificere de faktorer, der fører til potentielt skadelige beslutninger truffet af AI-modellerne, såsom fejl i data, fejlbehæftede algoritmer eller utilstrækkelige testprocedurer. For eksempel kan en AI-model fejlagtigt identificere civile som fjender på grund af bias i træningsdataene eller utilstrækkelig kontrol over algoritmens beslutningsprocesser. Derfor har dataloger og dataprofessionelle ansvar for at sikre, at dataene er repræsentative og uden bias, samt at algoritmerne testes grundigt i forskellige scenarier for at minimere risikoen for fejlagtige beslutninger. Deres kausale ansvar omfatter derfor de tekniske valg i udviklingen af AI-systemerne, herunder træning af modeller, valg af data og implementering af algoritmer, som kan påvirke systemets handlinger. Desuden bør de være opmærksomme på institutionelle faktorer som publikationspres og konkurrence om forskningsmidler, der kan påvirke kvaliteten af deres arbejde [1].

Moralsk ansvar i konteksten af LAWS betyder, at både de individuelle udviklere og de institutioner, der støtter udviklingen af disse systemer, skal holdes ansvarlige for de beslutninger og handlinger, som systemerne udfører. I denne sammenhæng ligger en stor del af ansvaret hos militæret, da det er dem, der i sidste ende anvender systemerne. Men ifølge Jesper Ryberg (2003) kan forskere og udviklere ikke undskylde sig med, at det er staten eller militæret, der bestemmer anvendelsen af deres teknologi. Ryberg understreger, at udvikleren har et medansvar for at overveje de potentielle anvendelser og konsekvenser af deres forskning. Hvis teknologien misbruges, bærer alle involverede et ansvar for at have muliggjort denne anvendelse. Det er ikke nok at hævde, at en anden ville have udviklet en lignende teknologi, hvis man selv havde undladt at gøre det. Ansvaret for at sikre, at teknologien anvendes etisk forsvarligt, skal deles mellem udviklerne, deres institutioner og de slutbrugere, som i dette tilfælde er militæret [2].
Denne ansvarsfordeling kræver, at udviklere aktivt vurderer de etiske implikationer af deres arbejde under hele udviklingsprocessen. Udviklerne bør være bevidste om de potentielle militære anvendelser af deres teknologi. Samtidig skal institutionerne implementere etiske retningslinjer og sikre, at der er mekanismer på plads til at overvåge og evaluere anvendelsen af teknologien. Dominante faglige foreninger som Institute of Electrical and Electronics Engineers (IEEE) har klare etiske regler, der klart tildeler medlemmerne af foreningerne et ret omfattende etisk ansvar. Ifølge IEEE's etiske regler skal medlemmerne bl.a. "to hold paramount the safety, health, and welfare of the public, to strive to comply with ethical design and sustainable development practices, and to disclose promptly factors that might endanger the public or the environment." [2]

Politisk ansvar i brugen af AI til udvikling af LAWS er afgørende for at sikre, at disse teknologier anvendes på en etisk og sikker måde. Regulatoriske rammer skal etableres og håndhæves for at sikre, at LAWS opererer inden for rammerne af international ret og menneskerettigheder, og for at der er mekanismer på plads for at holde de ansvarlige til regnskab. Det er ansvaret for politiske ledere at skabe internationale aftaler og samarbejde, som f.eks. mellem USA og Kina, hvor man har aftalt ikke at bruge AI til afsendelsen af atomvåben. I kontrast til dette har Rusland genoplivet et gammelt projekt, som kunne affyre atomvåben autonomt. Dataloger og dataprofessionelle har også et politisk ansvar for at samarbejde med beslutningstagere, bidrage til offentlig oplysning og nægte at deltage i udviklingen af systemer, der strider mod etiske standarder, såsom autonome atomvåben [6].

References

[1] S. Hilgartner, "Fraud, misconduct, and the irb," *IRB: Ethics & Human Research*, vol. 12, no. 1, 1990. [Online]. Available: https://www.jstor.org/stable/3563681

[2] H. K. Sørensen and M. W. Johansen, "Kapitel 10: Etik, redelighed og privacy," in *Invitation til de datalogiske fags videnskabsteori*, apr 2022.

[3] ——, "Kapitel 6: At få computeren til at hjælpe os: Del ii — datadrevne modeller," in *Invitation til de datalogiske fags videnskabsteori*, apr 2022.

[4] T. Miller, P. Howe, and L. Sonenberg, "Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences," *CoRR*, vol. abs/1712.00547, 2017. [Online]. Available: http://arxiv.org/abs/1712.00547

[5] I. Bode, "Falling under the radar: The problem of algorithmic bias and military applications of ai," *ICRC Humanitarian Law & Policy Blog*, 2024, accessed: 2024-06-06. [Online]. Available: https://blogs.icrc.org/law-and-policy/wp-content/uploads/sites/102/2024/03/falling-under-the-radar-the-problem-of-algorithmic-bias-and-military-applications-of-ai-PDF.pdf

[6] D. Adam, "Lethal ai weapons are here: How can we control them?" *Nature*, vol. 629, 2024. [Online]. Available: https://www.nature.com/articles/d41586-024-01029-0

[7] M. W. Johansen and H. K. Sørensen, "Big data's titanic?" *Aktuel Naturvidenskab*, vol. 3, 2018.