Hvad er Safe AI?

10. januar 2019

Af Lars Kai Hansen, DTU Compute,

Forskning i kunstig intelligens handler om at forstå muligheder og begrænsninger i computersystemer, der opsøger og virker i verden gennem aktiv sansning, læring og kommunikation.

Samfundsmæssigt er der enorme muligheder i kunstig intelligens. Computere kan samkøre meget mere information og de kan regne hurtigere og mere systematisk end mennesker. Derfor vil veludviklede intelligente systemer f.eks. gøre os bedre til at forstå, forebygge og behandle sygdomme, køre lastbiler hurtigere og mere sikkert gennem trafikken, og forudsige rækkevidden af politiske tiltag mere præcist. På baggrund af de samfundsmæssige konsekvenser, bliver der verden over arbejdet med at sikre at teknologien lever op til demokratiske værdier og kontrol[1].

På DTU Compute har vi formuleret en række *tekniske* 'Safe AI' principper[2]. Samlet udgør de en vision for ansvarlig kunstig intelligens, baseret på *konkret og realistisk AI-teknologi*:

Safe AI er sikker – har bestået test, er verificeret og robust overfor systematiske og velinformerede angreb.

Note: Test, validering og verifikation[3], er veldefinerede ingeniørvidenskabelige begreber. Verifikation sikrer at det intelligente system opfører sig som planlagt. Test henviser til at der er gennemført statistiske analyser af systemets virkning og begrænsninger. AI skal være certificeret robust overfor målrettede angreb[4] fra velinformerede modstandere.

Safe AI er selvbevidst – forstår sin egen rolle og usikkerhed, kan f.eks. afslå at handle.

Note: Bevidsthed er et veldefineret psykologisk begreb[5], som dækker over den koordinerede funktion i hjernen, der samler information fra sanser og kropsfunktioner. Målet med bevidsthed er at sikre at alle dele af systemet arbejder sammen om en given opgave og at der gives en rudimentær samlet beskrivelse af situationen, der kan kommunikeres til andre intelligente systemer. Herunder skal AI være bevidst om egen rolle, f.eks. egen usikkerhed, der kan danne baggrund for at nedlægge 'veto'[6], mod usikre handlinger.

---

[1] Se f.eks. The Partnership on AI. https://www.partnershiponai.org/about/
S. Pichai 2018 AI at Google: our principles https://blog.google/technology/ai/ai-principles/

[2] L.K. Hansen, 2018. Kunstig intelligens – dommedag eller Safe AI nu? Mandag Morgen 24 juni. Link: https://www.mm.dk/artikel/kunstig-intelligens-dommedag-eller-safe-ai-nu

[3] Pham, H. (1999). Software Reliability. John Wiley & Sons, Inc.

[4] Se f.eks.: Kurakin, A., Goodfellow, I. and Bengio, S., 2016. Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.

[5] Dehaene, S., et al., 2017. What is consciousness, and could machines have it?. Science, 358(6362), pp.486-492.

[6] Se f.eks.: Hansen, LK, et al., 1997. The error-reject tradeoff. Open Systems & Information Dynamics, 4(2), pp.159-184.

Safe AI kan holde på en hemmelighed – har indbygget beskyttelse af privatliv, 'privacy by design'.

Note: Kendskab til private oplysninger kan misbruges til at styre adfærd og forbrug - f.eks. til at opnå urimelige konkurrencefordele eller til politisk styring. Intelligente systemer kan designes og implementeres så de respekterer privatlivsgrænser[7]*. Med såkaldt 'differential privacy' er det teknisk muligt at lære værdifulde generelle sammenhænge - uden at man kompromitterer enkeltpersoners private oplysninger*[8]*.*

Safe AI har veldefinerede værdier – er renset for stereotyper, 'bias', og forstår emotioner

Note: Man kan sikre at et intelligent systems beslutninger sker i overensstemmelse med et specifikt værdisæt. Man kan for eksempel undgå at AI tager beslutninger, der baserer sig på fordomme[9] *eller på uønskede stereotyper*[10]*. Det er teknisk muligt for AI at opfatte, reagere på og kommunikere med følelser*[11]*.*

Safe AI har sociale kompetencer – forstår sociale relationer, forstår brugerens viden og kompetencer.

Note: Samfundsmæssige fremskridt afspejler den gensidige afhængighed der opstår mellem mennesker ved specialisering, arbejdsdeling og kommunikation. Det fremhæves ofte at det danske samfund bygger på en høj grad af tillid til medborgere, virksomheder og det offentlige. Det er vigtigt at kunstig intelligens forstår sine omgivelser og brugere, forstår deres relationer[12]*, viden, muligheder og begrænsninger*[13]*, og kan kommunikere effektivt og tillidsvækkende.*

Safe AI forstår magt – forstår data og handlingers kontekst og konsekvens

Note: Viden er magtens grundlag. Vidensgrafer er en teknisk betegnelse for de hjernefunktioner, som vi hos mennesket kalder hukommelse, paratviden, sund fornuft, erfaring, intuition osv. Google bruger f.eks. sin vidensgraf til at fortolke søgninger og kommunikere resultater. Vidensgrafer kan være generelle, som Wikipedia, eller specifikke og indeholde viden om en bestemt person eller situation. Vidensgrafer kan hjælpe til at navigere i magtforhold, dvs. et intelligent systems kontekst og forståelse af sine handlingers konsekvens[14]*. Vidensgrafer er mere end statistik, de repræsenterer fysisk forståelse for årsag og virkning*[15]*.*

---

[7] Se f.eks. https://dataethics.eu/

[8] Se f.eks.: Dwork, C., et al., 2006, Calibrating noise to sensitivity in private data analysis. In Theory of Cryptography Conference. Springer, Heidelberg

[9] Se f.eks.: Agarwal, A., et al. 2018. A reductions approach to fair classification. arXiv preprint arXiv:1803.02453.

[10] Se f.eks.: Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V. and Kalai, A.T., 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in Neural Information Processing Systems (pp. 4349-4357).

[11] Se f.eks.: Nielsen FÅ, 2011 A new ANEW: evaluation of a word list for sentiment analysis in microblogs Proc ESWC2011 Workshop p. 93-98.

[12] Se f.eks. Mislove, A., Lehmann, S., Ahn, Y.Y., Onnela, J.P. and Rosenquist, J.N., 2011. Understanding the Demographics of Twitter Users. ICWSM, 11(5th), p.25.

[13] Definition: https://www.google.com/intl/bn/insidesearch/features/search/knowledge.html.

[14] Se f.eks.: Moravčík, M., et al. 2017. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. Science, 356(6337), pp.508-513.

[15] Pearl, J. and Mackenzie, D., 2018. The Book of Why: The New Science of Cause and Effect. Basic Books.

Safe AI er dokumenteret – transparent, kommunikerende, ”right to explanation”.

Note: Modsat den generelle opfattelse, er det i mange tilfælde teknisk muligt at forklare beslutninger der er taget af kunstig intelligens. Det er muligt at skabe transparent AI og forklare både den generelle virkning af et AI modul[16] og baggrunden for specifikke beslutninger. Det er derfor muligt at skabe AI systemer der kan leve op til de demokratiske forventninger, der er nedfældet i de europæiske GDRP regler[17]. Det er teknisk muligt at revidere et kommercielt, eller offentligt, AI systems metoder og virkning, på samme måde som samfundet i dag kræver adgang og godkendelse af en virksomheds økonomiske og miljømæssige regnskaber.

Safe AI er *”open source”* – metoder, kode og testresultater er tilgængelige

Note: Demokratisk kontrol med intelligente systemer kan forbedres gennem transparens og dyb indsigt i design, implementering og test resultater. Der er ingen modsætning mellem forretning og open source[18]. Open source kan hjælpe til at reducere fejl og mangler i et system. Open source skaber tillid.

[16] Se f.eks.: Vilamala, A., et al., 2017. Deep Convolutional Neural Networks for Interpretable Analysis of EEG Sleep Stage Scoring. arXiv preprint arXiv:1710.00633.
[17] Se f.eks.: Ribeiro, M.T. et al 2016, Why should i trust you?: Explaining the predictions of any classifier. In Proc 22nd ACM SIGKDD.
[18] Se f.eks.: F.eks. https://www.tensorflow.org/