AI502

1. INTRODUKTION TIL DEN KUNSTIGE INTELLIGENS' ETIK

AI502 : KURSETS FORMAT

Format: 13x2 timers undervisning, 5 ECTS

6x2 timer primært om etik, 7x2 timer primært om jura

Undervisere:

Nikolaj Nottelmann, filosofi

Ayo Næsborg-Andersen, jura

Jakub Skórczynski, jura

Afsluttende mundtlig eksamen i januar, 20 min inkl. votering, bedømmelse: B/IB

Liste over eksamens-cases udleveres på forhånd efter undervisningens afslutning

AI 502: KURSETS INDHOLD

I forhold til uddannelsens kompetenceprofil har kurset eksplicit fokus på at:

give kompetencer til at håndtere arbejds- og udviklingssituationer, således at spørgsmål om etik og privathed inddrages pro-aktivt i designfasen.

give færdigheder i analyse af og refleksion over it-etiske problemstillinger, herunder brug af metoder, såsom f.eks., Privacy by Design.

give viden om etiske problemstillinger ifm udvikling og anvendelse af kunstig intelligens, fx vedrørende bias i algoritmer, transparens, fairness og profilering.

give viden om teorier om privathed og privathedsfremmende designmetoder, såsom f.eks. Privacy by Design.

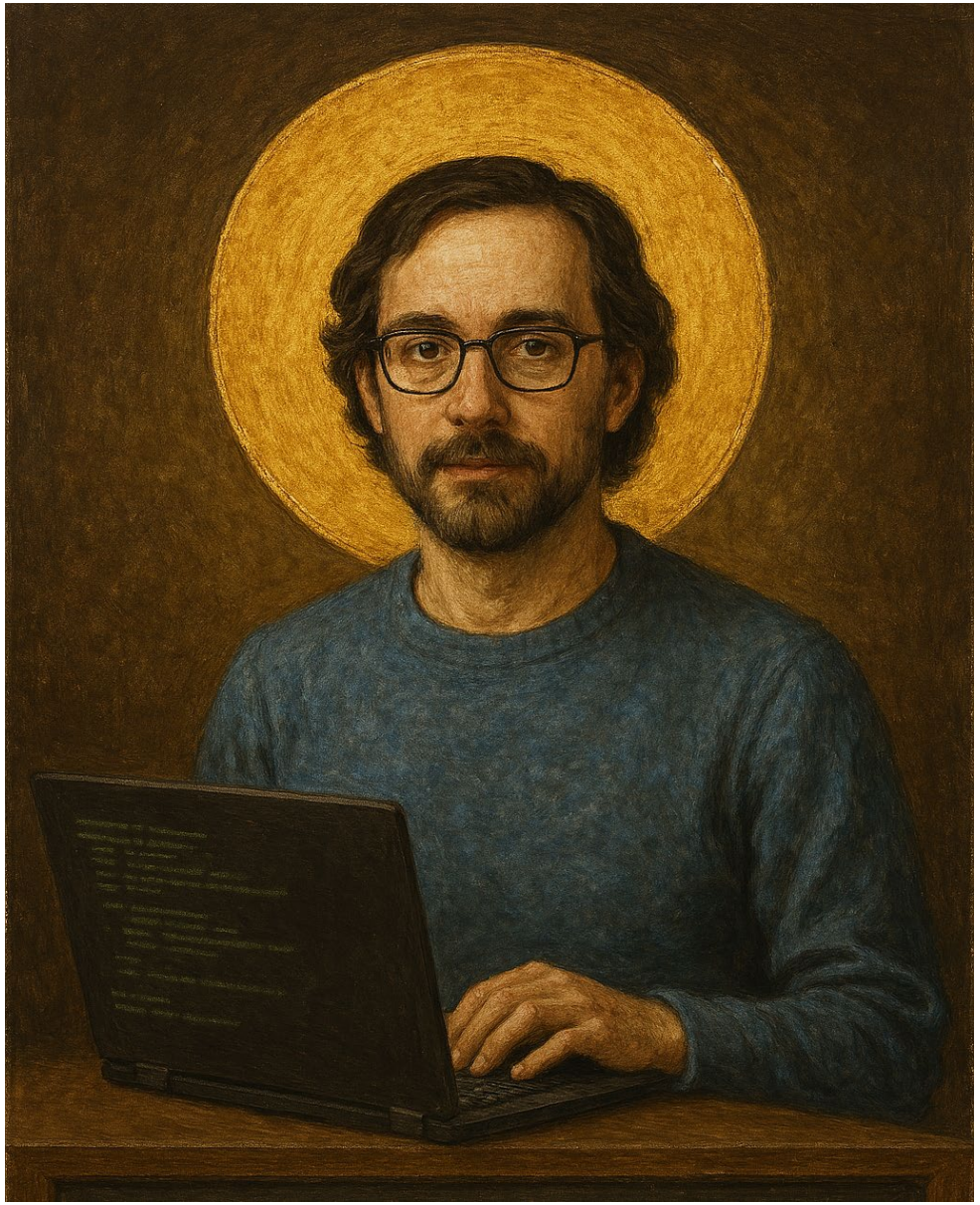
give viden om grundlæggende retlige problemstillinger vedrørende kunstig intelligens

HVORFOR SKAL I LÆRE NOGET OM ETIK?

Samfundet og arbejdsmarkedet har brug for (og bør bruge) it-professionelle og datalogiske teoretikere, der kan:

1. Tænke etisk forsvarlige løsninger ind i både design –, udviklings- og implementeringsfaser

2. Bidrage kompetent til organisations-interne og offentlige debatter om etiske spørgsmål ifm. AI



DEN ETISKE DEL AF AI502 - TEMAER

1. Introduktion : de grundlæggende AI-etiske begreber og problemstillinger
2. AI som en eksistentiel trussel mod menneskeheden? Hvorfor/hvorfor ikke?
3. Grundlæggende maskinetik og maskinmetaetik
4. Etisk AI i sundhedsvæsen og militær
5. Etiske problemer vedr. AI og masseovervågning
6. Etiske problemer vedr. AI og paternalistisk lovgivning

HVAD HANDLER ETIK OM?

Forståelse og løsning af etiske problemer !

Et etisk problem = Et problem om korrekt afvejning af værdier

Udfordring: Værdier kan afvejes ud fra mange hensyn: Hvad er mest praktisk? Hvad er smukkest osv.

Svar: **Den korrekte afvejning af værdier er den, som vi, alt taget i betragtning, bør foretage i den relevante situation**

Giv et eksempel på et etisk problem vedr. AI!

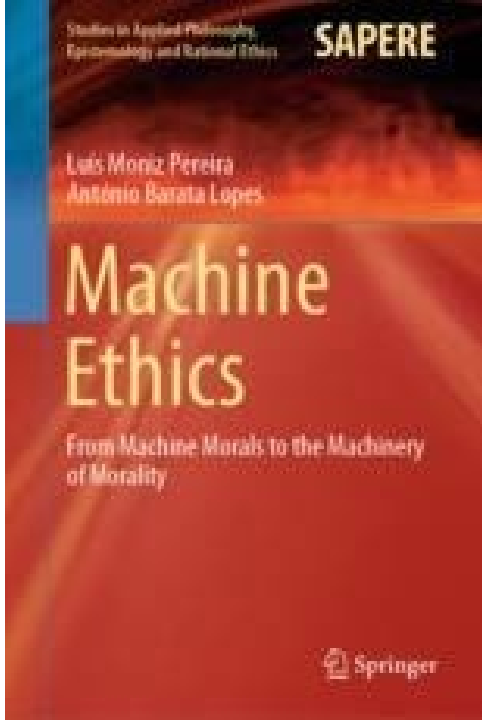
2

Hhh

Hvad er maskinetik?

→ we restrict the term “machine ethics” to research which directly contributes to the creation of ethical machines. This includes attempts by engineers and scientists to actually build such machines and theoretical research aiming to facilitate or enable this, but not broader philosophical inquiries into the implications of this technology. The latter field…is sometimes called “machine metaethics”

→ Cave et al. 2019, 562



Maskinetik := Læren om konstruktion af etiske robotter

Maskin**meta**etik := Læren om bredere problemer (fx begrebslige, sociale, politiske) ifm. udvikling og brug af etiske robotter

FINDES DER OVERHOVEDET ETISK KORREKTE LØSNINGER?

Uanset om der findes korrekte løsninger på nogle etiske problemer:

1. Det kan være meget svært at begrunde et svar på et etisk problem!

2. Der er dyb og hårdnakket uenighed om nogle etiske problemer, fx fri abort og aktiv dødshjælp

3. Nogle etiske problemer er måske uløselige

NB! 1-3 betyder IKKE at INGEN etiske problemer har korrekte svar; eller at det ikke er umagen værd at lede efter korrekte svar på ETHVERT etisk problem!

GRUNDLÆGGENDE ANALYSE AF ET ETISK PROBLEM

1. VÆRDIER: Hvilke **værdier** er på spil?

2. MORALSK AGENT: Hvem har **ansvaret** for at vælge rigtigt?

3. MORALSKE PATIENTER: Hvem skal der **tages hensyn til**?

MORALSK STATUS – NÆRMERE OM MORALSKE AGENTER OG PATIENTER

En **moralsk agent**

1. Kan handle
2. Er underlagt forpligtelser, således at hun kan fortjene bebrejdelser eller ros for sine handlinger
   - Moral agents are portrayed by the traditional moral theories as having certain obligations. It is not just that it would be nice or a good thing if [she] acted in certain ways (Nyholm 2021, 49)

En **moralsk patient**

1. Kan lide skade
2. Kan krænkes/forurettes
   - [is] somebody against whom we can act rightly or wrongly (Nyholm 2021, 45)

MORALSKE PATIENTER, DER IKKE ER MORALSKE AGENTER

Små børn

Dyr med bevidsthed?

Nogle robotter? [mere om dette senere]

Kan der findes en moralsk agent som ikke samtidig er en moralsk patient? Begrund dit svar!

Nobody has responded yet.

Hang tight! Responses are coming in.

DET SKYLDFRI VALG VS.
DET RIGTIGE VALG VS.
DET ROSVÆRDIGE VALG

IKKE helt det samme!

1. En moralsk agent kan være undskyldt (derfor skyldfri) for at vælge forkert

2. En moralsk agent kan vælge det rigtige uden at fortjene ros, fordi hun blot har gjort sin minimale pligt

3. Mere kontroversielt: En agent kan være skyldig for at vælge at gøre det rigtige, hvis hun vælger det af helt forkerte grunde!

Kan en agent være rosværdig for en moralsk forkert handling? Begrund dit svar!

0

Nobody has responded yet.

Hang tight! Responses are coming in.

VÆRDITEORI:
TEORIER OM RELEVANT VÆRDI FOR ETISKE PROBLEMER

1. Iflg KONSEKVENTIALISTER har kun FREMTIDIGE UDFALD en etisk relevant værdi
   - fortiden og nutiden er det alligevel for sent at ændre på!

2. Iflg. nogle PLIGTETISKE teorier har også NUTIDIGE og FORTIDIGE rettigheder og fortjenester relevant værdi
   - Alle personer, også fortidige og nutidige, har en værdighed som kan krænkes
   - Det krænker personers værdighed ikke at give dem som fortjent eller som de har ret til
   - Det har negativ etisk værdi at krænke personers rettigheder, **uanset udfaldet**

Antag at vi drastisk kan formindske rygerrelaterede onder ved årligt at henrette fem tilfældige rygere. Hvem ville være gladest for denne løsning?

Konsekvensetikere

0%

Pligtetikere

0%

Begge positioner er forpligtet på at afvise løsningen som etisk forkert

0%

I PRAKSIS OFTE ENIGHED MELLEM FREMADSKUENDE OG TILBAGESKUENDE TEORIER

Skal jeg holde mit løfte til mormor om at pleje hendes gravsted eller nyde tiden med min Playstation?

Konsekventialistiske og historiske teorier kan begge svare DET FØRSTE!! men med forskellige begrundelser:

1. Konsekventialisten kan fx pege på at jeg sætter et dårligt eksempel for andre, hvis jeg svigter mit løfte. At mormors gravsted forfalder osv.
2. Den tilbageskuende teori kan pege på, at mormor har **fortjent** at jeg holder mit løfte – uanset hvor surt det er for mig og andre

FINDES DER ULTIMATIVE VÆRDIER?

En ultimativ/fundamental/kategorisk værdi: Noget værdifuldt, i kraft af hvilket **alt andet** værdifuldt har værdi

Filosoffer har givet mange bud på, hvad der har ultimativ værdi, fx

1. Velfærd (at det går mennesker og evt. andre moralske patienter godt)
2. Lykkefølelse
3. Økologisk balance
4. At moralske agenter udvikler og udtrykker en dydig personlighed
5. At moralske patienters rettigheder bliver respekteret
6. Handlinger udført i respekt for en fornuftsgiven Morallov

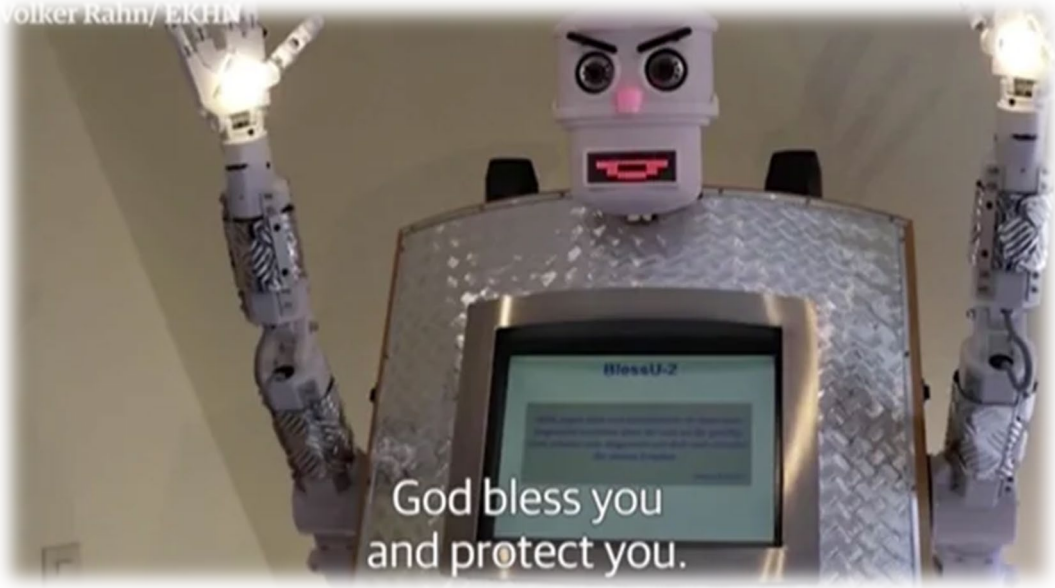
ER DET VIGTIGT OM DER FINDES ULTIMATIVE VÆRDIER?

I praksis oftest NEJ!

1. Vi kan diskutere mange etiske problemer relateret til AI uden antagelser om ultimative værdier

2. Vi kan godt stræbe efter (delvis) enighed om afvejning af værdier, uden at kunne sætte alle værdier på én fællesnævner

ROBOTTER SOM MORALSKE AGENTER OG PATIENTER



Hvad skal bæstet hedde?

Tilbage til Dartmouth College 1956!

Herbert Simon, Allen Newell: **Kompleks informationsprocessering?**

John McCarthy, Marvin Minsky: **Kunstig Intelligens!**

Problem 1: ”Kunstig” som i ”kunstig diamant” (som er diamant) ellers som i ”kunstige blomster” (som ikke er blomster) ?

Problem 2: Intelligens → Bevidsthed → Moralsk status?!

OPERATIONALISME MHT. INTELLIGENS

”Andeprincippet”: If is walks like a duck and quacks like a duck, it is a duck!!

M.a.o.: Hvis et system opfører sig intelligent, så ER det intelligent!

Se især Turing (1950)

Men medfører intelligens så bevidsthed? Eller moralsk status?

ONDSKAB MOD EN ROBOT?

https://youtu.be/0VgxAnZKM14?si=LJK1o0LiCZP2Hjs5

Er ATLAS robotten i filmen en moralsk patient? Begrund dit svar!

Nobody has responded yet.

Hang tight! Responses are coming in.

HVAD HVIS ROBOTTER FORELØBIG **IKKE** HAR MORALSK STATUS?

1. Stor lettelse ift. færre moralsk patienter: Må vi slukke for en moralsk patient?

2. Gode modeller for, hvorfor de alligevel nogle gange bør behandles som om de var moralske patienter (fx forbud mod pædofil billedfiktion)

3. Den moralske agentstatus (”aben”) falder tilbage på robottens designere, producenter, programmører og brugere

OPTAKT TIL GRUPPEOPGAVE: DEN NATIONALE STRATEGI FOR KUNSTIG INTELLIGENS (ERHVERVS- OG FINANSMINISTERIET 2019)

Nogle positive værdier/gevinster (s. 11)

1. Mere, hurtigere, og bedre hjælp og behandling til borgere
2. Bedre informationssøgning
3. Udvikling af nye forretningsmodeller
4. Afdækning af administrative fejl
5. Afsløring af lovovertrædelser
6. Overvågning af systemer og miljø

Nogle måske truede værdier (s. 28-9)

1. Menneskers selvbestemmelse
2. Menneskers værdighed
3. At nogen kan stilles til ansvar for skader
4. At borgerne kan forstå offentlig sagsbehandling
5. At ingen diskrimineres uretfærdigt pga. fordomme
6. Fremskridt i form af bedre offentlig service og økonomisk vækst

Giv et eksempel på et etisk problem hvor gevinsterne ved AI er i konflikt med værdien i menneskelig selvbestemmelse!

Nobody has responded yet.

Hang tight! Responses are coming in.

Giv er eksempel på et etisk problem hvor gevinsterne ved AI er i konflikt med værdien i at kunne stille nogen til ansvar for skader!

Nobody has responded yet.

Hang tight! Responses are coming in.

Giv et eksempel på et etisk problem hvor gevinster ved AI er i konflikt med værdien i at ingen diskrimineres uretfærdigt pga. fordomme!

Nobody has responded yet.

Hang tight! Responses are coming in.