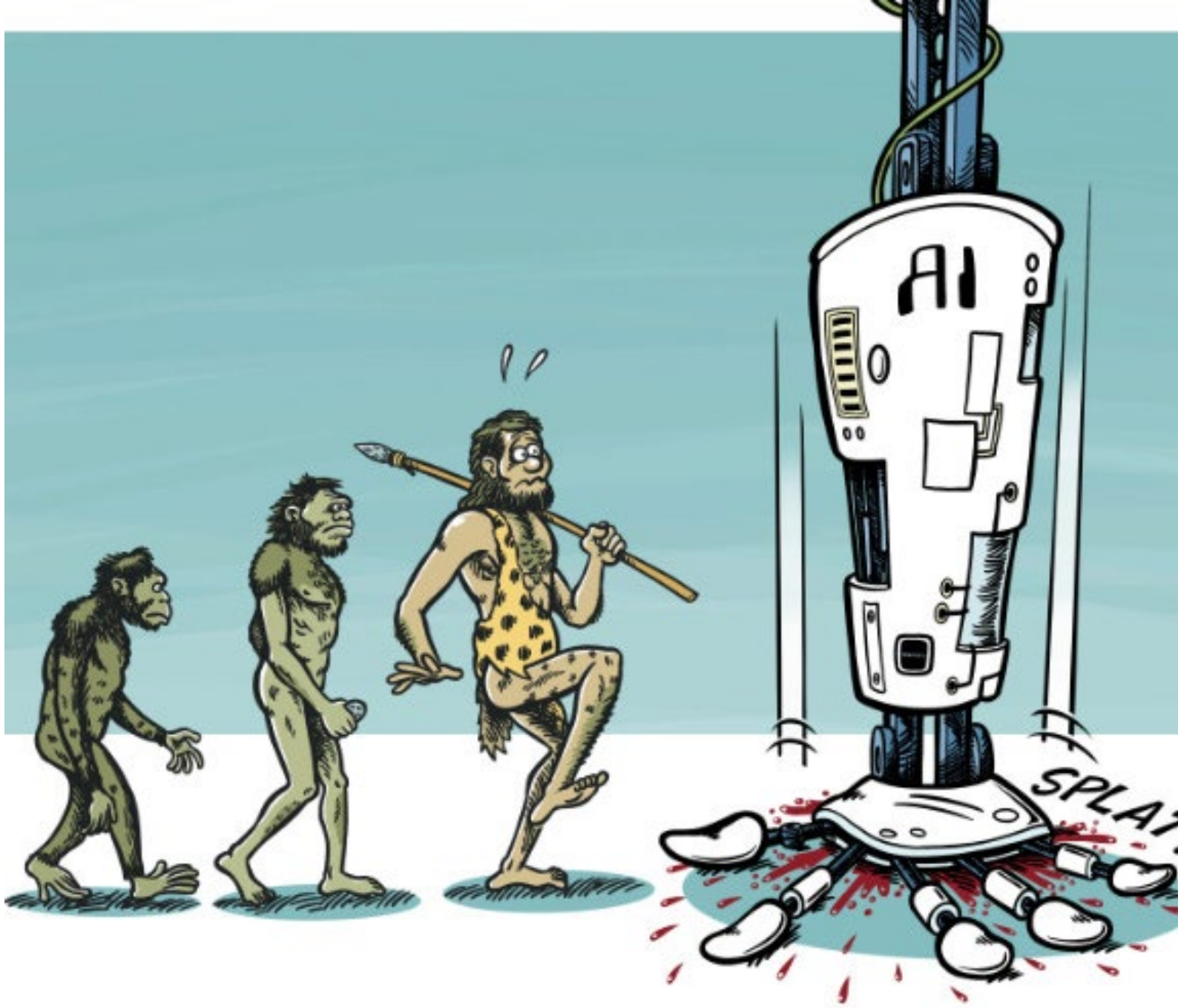
AI502

2. UDGØR KUNSTIG INTELLIGENS EN EKSISTENTIEL TRUSSEL?



REPETITION:
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**

REPETITION: GRUNDLÆGGENDE ANALYSE AF ET ETISK PROBLEM

1. VÆRDIER: Hvilke **værdier** skal afvejes?
2. MORALSK AGENT: Hvem har **ansvaret** for at vælge rigtigt?
3. MORALSKE PATIENTER: Hvem skal der **tages hensyn til**?

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

Nogle positive værdier/gevinster (s. 11)

1. Mere, hurtigere, og mere individuel 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 skadevolden
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.

AI SOM EKSISTENTIEL TRUSSEL – EN ALARMISTISK DIAGNOSE

Once criticized in Bloomberg for being an AI "doomer," Yudkowsky says he's not the only person "steeped in these issues" who believes that "the most likely result of building a superhumanly smart AI, under anything remotely like the current circumstances, is that literally everyone on Earth will die." He has the receipts to back it up, too, citing an expert survey in which a bunch of the respondents were deeply concerned about the "existential risks" posed by AI. These risks aren't, Yudkowsky wrote in Time, just remote possibilities….There is, to Yudkowsky's mind, but one solution to the impending existential threat of a "hostile" superhuman AGI: "just shut it all down," by any means necessary

- https://futurism.com/ai-expert-bomb-datacenters

KLASSIFIKATION AF EN TRUSSEL

1. Omfang (*Scope*) : Hvor **mange** etiske patienter berøres negativt?
2. Sværhedsgrad (*Severity*): Hvor **meget** påvirkes disse patienter som minimum?
3. Sandsynlighed (*Risk*): Hvor **sandsynligt** er scenariet?

HVAD ER EN **EKSISTENTIEL** TRUSSEL (*XRISK*)?

Xrisks are at the most extreme end of both of these spectrums: they are pan-generational in scope (i.e. 'affecting humanity over all, or almost all, future generations') and they are the severest kinds of threats, causing either 'death or a permanent and drastic reduction of quality of life' (Bostrom 2013: 17).

- Vold & Harris 2021, 725-6

1. **Omfang**: Påvirker (næsten) **alle mennesker i alle fremtidige generationer**

2. **Sværhedsgrad**: Medfører **død** eller - i bedste fald - **drastisk reduceret livskvalitet** ← Vore en vasoft risiko

3. **Sandsynlighed** (NB! ikke nævnt i citatet): **væsentlig!**

Beskriv en AI-relateret Xrisk i Bostrom's forståelse (ekstremt omfang og sværhedsgrad, se bort fra sandsynlighed)!

Nobody has responded yet.

Hang tight! Responses are coming in.

HVAD ER EN "VÆSENTLIG" SANDSYNLIGHED?

Sammenlign flg. hypotetiske scenarier:

+ F

1. Vi ved (har stærk grund til at tro) at AI udløser en pangenerationel katastrofe, medmindre vi griber ind (som påstået af fx Yudkowsky)
   - Her bør vi oplagt gribe ind (jf Yudkowsky)!
2. Vi ved at AI med mindst X % sandsynlighed (X>>0) vil udløse en pangenerationel katastrofe, medmindre vi griber ind
   - Her afhænger den korrekte strategi af, hvor **risikovillige** vi bør være i lyset af goderne ved **ikke** at gribe ind, fx mindskelsen af **andre** eksistentielle risici!
3. Vi kan ikke udelukke, at AI vil udløse en pangenerationel katastrofe, medmindre vi griber ind
   - En blot "risiko for en risiko" - Dette bør næppe bekymre os i sig selv!

XRISK SOM ETISK TRUMF

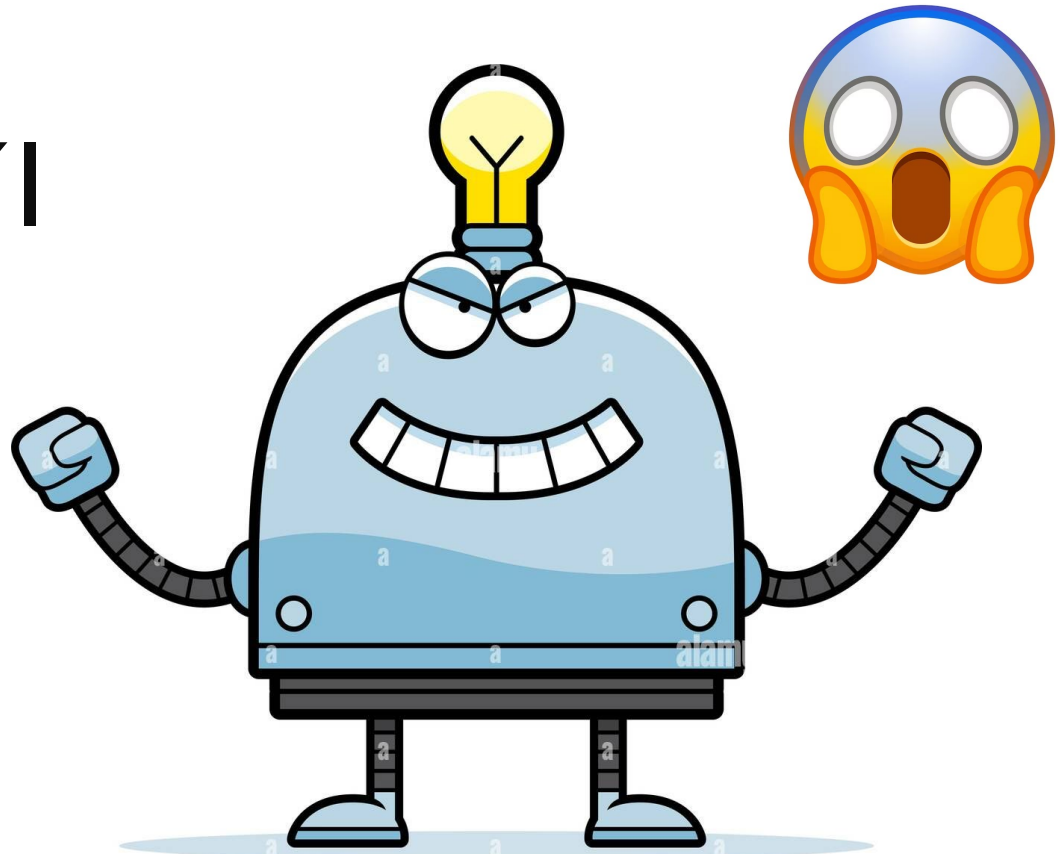
Antag:

1. En moralsk agent A har valget mellem udfaldene U og W
2. Kun ét af udfaldene U og W indebærer en **nogenlunde sandsynlig eksistentiel trussel**

Så bør A <u>fravælge den eksistentielle trussel</u>, uanset øvrige omstændigheder!

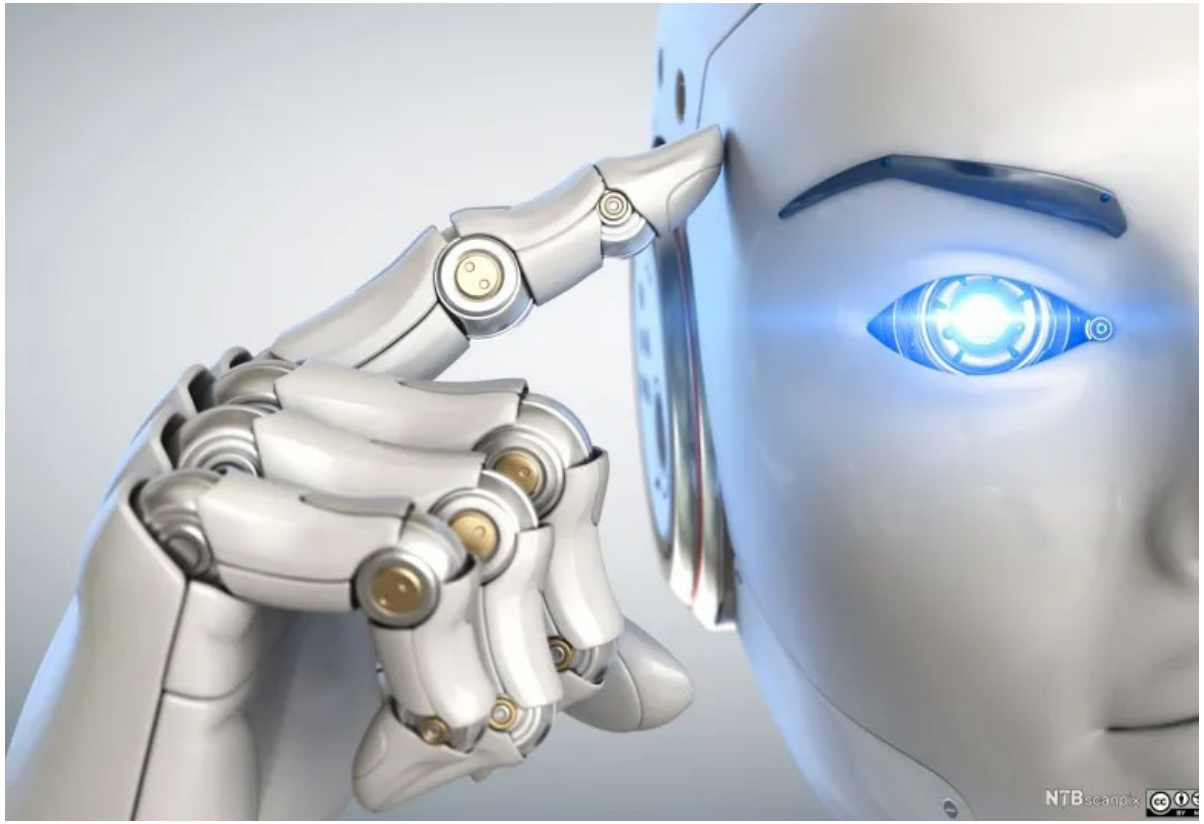
Jf. Yudkowsky: **Bedre (U) at vi bomber alle servere end (W) at AI snart udsletter os!**

TYPER AF AI-RELATERET *XRISK* I TERMER AF SVÆRHEDSGRAD



1. Menneskehedens udslettelse
   - Vi bliver udryddet af robotter
   - Vi uddør pga. ressourcemangel
     - Robotterne bruger ressourcerne
     - Robotterne hamstrer ressourcerne
     - Robotterne ødelægger ressourcerne

2. Drastisk og irreversibelt tab af værdier, der gør menneskelivet værd at leve
   - I bedste fald et ekstremt lidelsesfuldt liv for overlevende mennesker
   - Civilisatorisk sammenbrud
   - Tab af moralsk dømmekraft og menneskelig værdighed

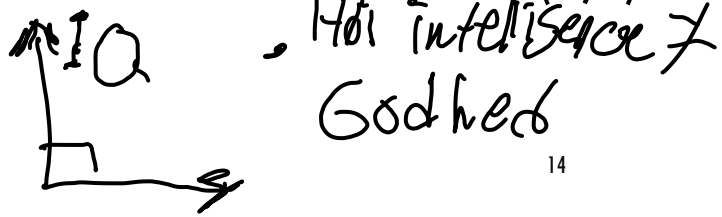


FORSKELLIGE MÅDER AI KAN MEDFØRE RISICI (JF. ZWETSLOOT AND DAFOE 2019, CIT. VOLD & HARRIS 2021, 727)

1. **Accidentielle risici:** Risiko for at AI-systemer uventet opfører sig på skadelige måder
2. **Strukturelle risici:** Indirekte risici som følge af, hvordan brugen af AI påvirker sociale og fysiske strukturer
3. **Misbrugsrisici:** Risiko for at ondsindede agenter bruger AI-systemer til at gøre skade

KONTROLPROBLEM-ARGUMENTET FOR *XRISK*
(= *CPAX*, V&H 2021, 727-8; CF. BOSTROM 2012)

1. Det er muligt at AI-systemer udvikler sig til at være menneskelig intelligens overlegent
2. Hvis et AI-system er menneskelig intelligens overlegent, så er det muligt, at vi ikke kan kontrollere det
3. Ergo: Det er muligt at ukontrollerbare AI-systemer udvikler sig (1,2)
4. **Ortogonalitetstesen:** Et AI-systems overlegenhed i intelligens er uafhængigt at dets målsætninger
5. Ergo: Det er muligt at ukontrollerbare AI-systemer udvikler sig med målsætninger, der ikke harmonerer med menneskelige målsætninger (3,4)
6. **Instrumentel Konvergens:** Et AI-system med målsætninger der ikke harmonerer med menneskelige målsætninger, vil tendere til at anvende midler, der undergraver menneskehedens overlevelse
7. Ergo: **Det er muligt at AI-systemer udvikler sig til en eksistentiel trussel mod menneskeheden!** (5,6)



Hvilken type risiko påpeger kontrolproblemet på ?

Accidentiel risiko 0%

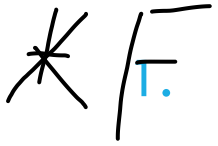
Strukturel risiko 0%

Misbrugsrisiko 0%

En helt anden type risiko 0%

UNDERARGUMENT FOR INSTRUMENTEL KONVERGENS

In Omohundro's view, '[a]ll computation and physical action requires the physical resources of space, time, matter, and free energy', and hence, 'almost any goal can be better accomplished by having more of these resources' (2008: 491). Bostrom (2014: 114–116) argues that, for this reason, it is likely that 'an extremely wide range of possible final goals' would generate 'the instrumental goal of unlimited resource acquisition (V&H 2021, 735)



1. Uanset sine strategiske målsætninger, vil et magtfuldt AI-system tendere til at samle og anvende væsentlige ressourcer
2. Hvis et AI-systems målsætninger ikke harmonerer med menneskehedens, vil det se nogle af menneskehedens nødvendige ressourcer som *fair game*
3. ERGO: Et magtfuldt AI-system med målsætninger der ikke harmonerer med menneskelige målsætninger, vil tendere til at anvende midler, der undergraver menneskehedens overlevelse (1,2)

EKSPLOSION I KUNSTIG INTELLIGENS?

Vil AI-systemer (selv)udvikle sig til faretruende super-intelligenser?

- *it is difficult, perhaps impossible, to predict what the motivations of a future advanced AI system could be. Ultimately, an intelligence explosion is certainly not an inevitable outcome, but it also is not an impossible one (V&H 2021, 720)*

Men farlig ukontrollerbarhed **kræver** måske slet ikke super-intelligens? (V&H, 2021, 732)

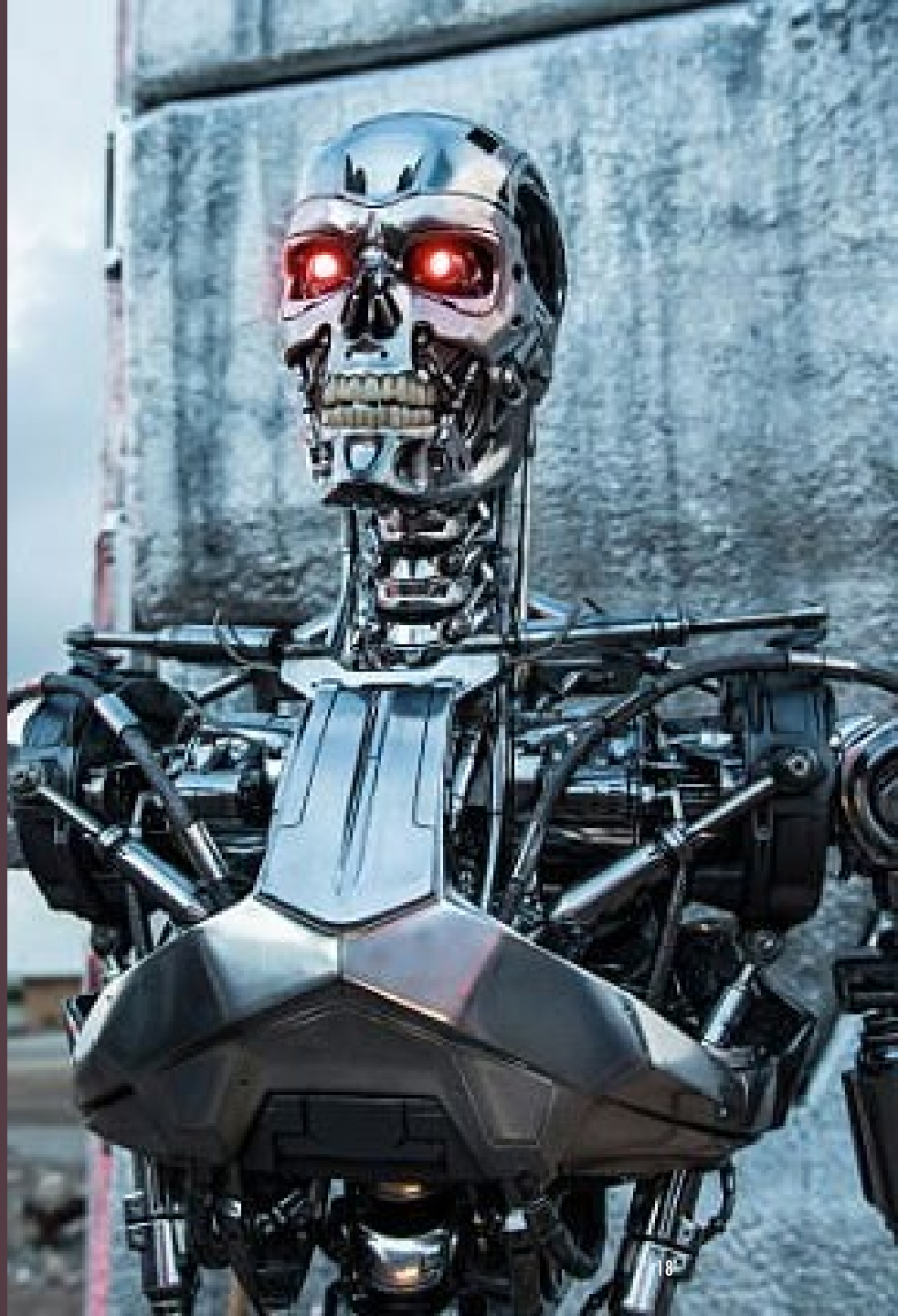
Også en sub-super-intelligent AI risikerer at anse mennesker for "myrer" der kan ofres for vigtigere mål!

CPAX'S MORALE: BYG KUN ETISKE ROBOTTER!

Undgå at AI-systemer bliver faretruende ukontrollerbare!!

Dette kan (forhåbentlig) opnås ved en **kombination** af:

1. Giv ikke robotter magt over destruktive midler!
2. Undgå at robotter bliver i stand til accelereret evolution (AI-eksplosion)!
3. Byg etiske *safeguards* ind i AI (= **etiske robotter**)!



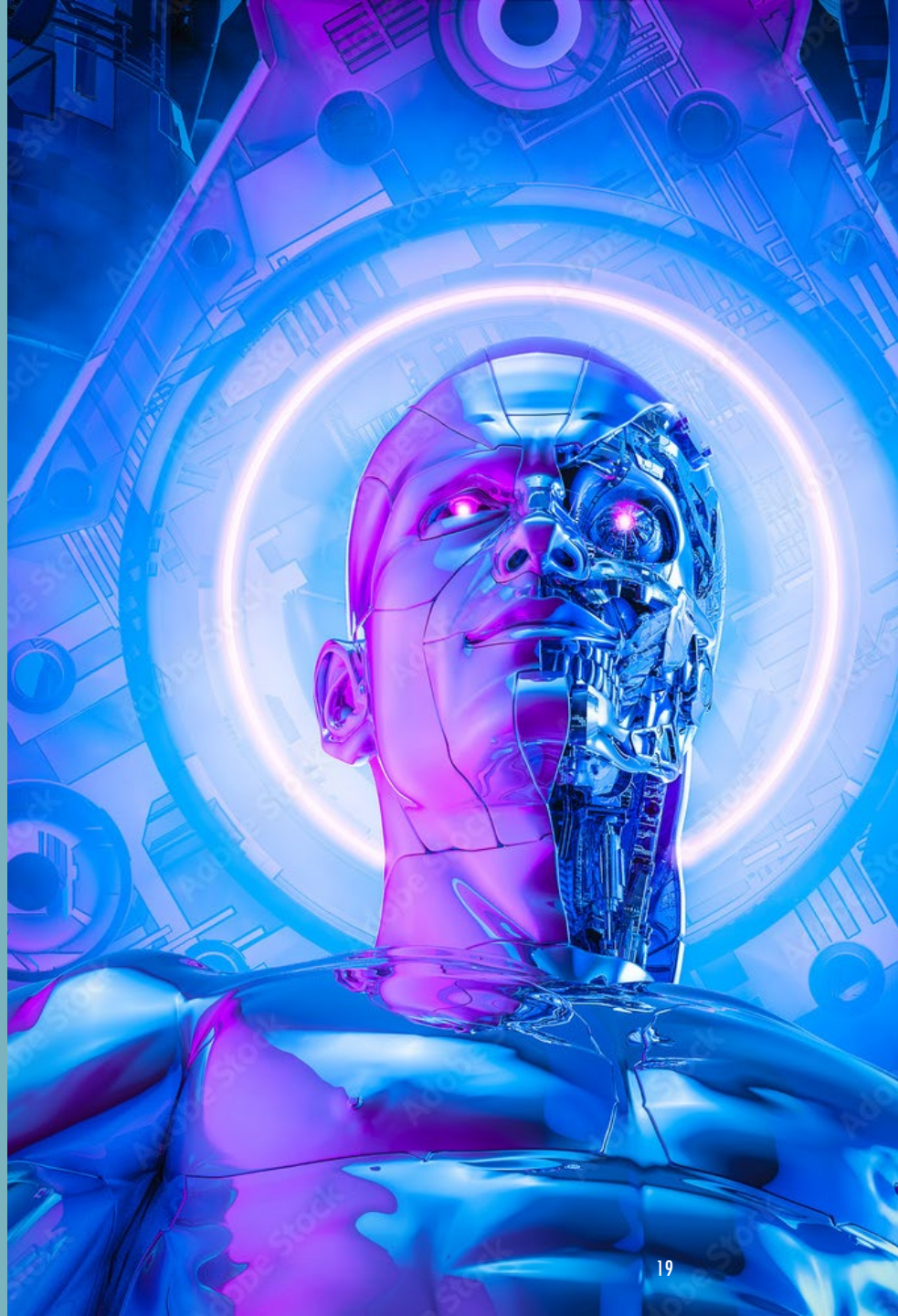
DEN ETISKE ROBOT

En etisk robot := En robot der (som minimum) pålideligt simulerer en etisk kompetent moralsk agent inden for sit arbejdsfelt

1. Er det godt nok at simulere et etisk kompetent **menneske**?
   - *The Problem of Human Moral Imperfection* (V&H 2021, 734)

2. **Hvis** forestilling om en etisk kompetent moralsk agent skal robotten simulere?
   - *Human beings are often confused and conflicted about our own values, and different cultures seem to have wide variation between their respective values* (ibid.)

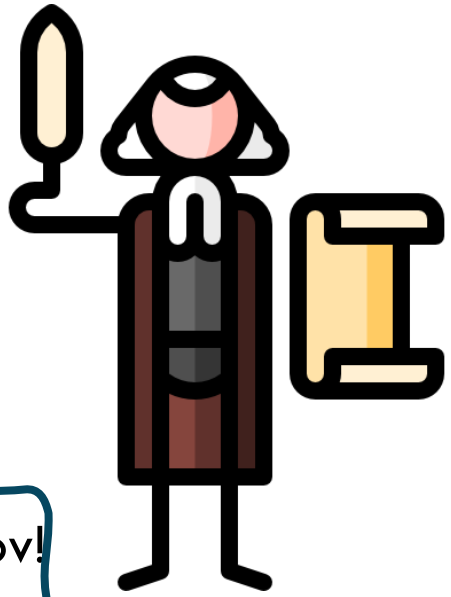
3. Hvordan gør vi plads til **moralske fremskridt**?
   - *Even if we can find a way to build machines that align with (only) the better parts of our current values, we would not want AI systems to codify these values in a way that prevents moral progress* (ibid.)



KANTS "KATEGORISKE IMPERATIV" (DVS. UBETINGEDE PÅBUD)

Immanuel Kant (1724-1804) påstod at have opdaget en moralsk regel der

1. Gælder undtagelsesløst
2. Altid har højeste prioritet
3. Kan begrunde alle andre moralske regler



Handl kun efter grundsætninger, som du samtidig kan ville ophøjes til almen lov!

Iflg. Kant analogt med:

Handl således at du altid betragter menneskeligheden hos både dig selv og andre som et mål i sig selv, ikke kun som et middel!

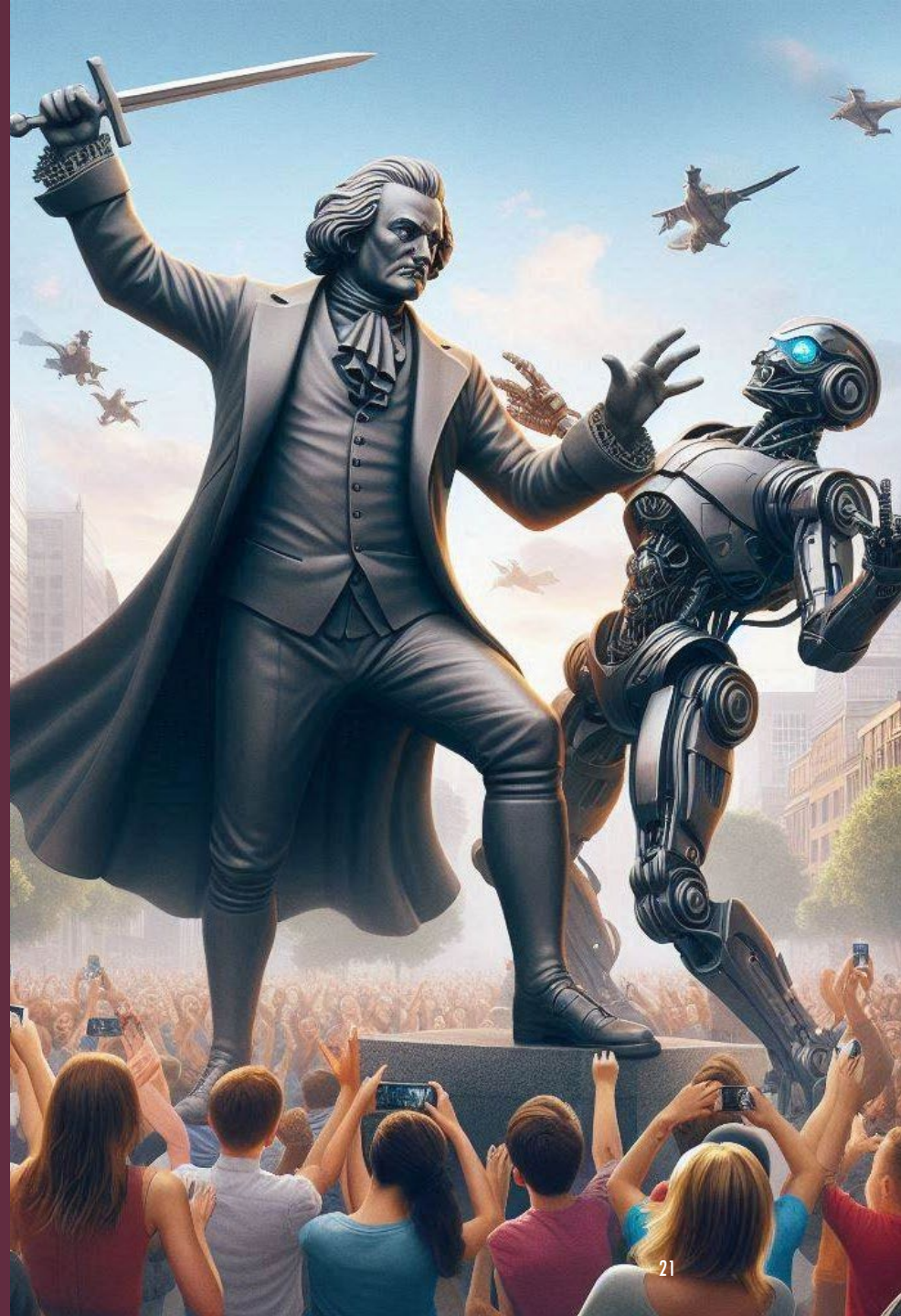
KRITIK AF KANT

For at gøre sin regel **universel**, gør Kant den samtidig **fuldstændig abstrakt**

Uden henvisning til **konkrete** moralske agenter, patienter, eller situationer

Dette gør den **ubrugelig** i praksis

Ikke mindst for **robotter** som ikke kan anlægge en menneskelig fortolkning!



"MIDAS-PROBLEMET"

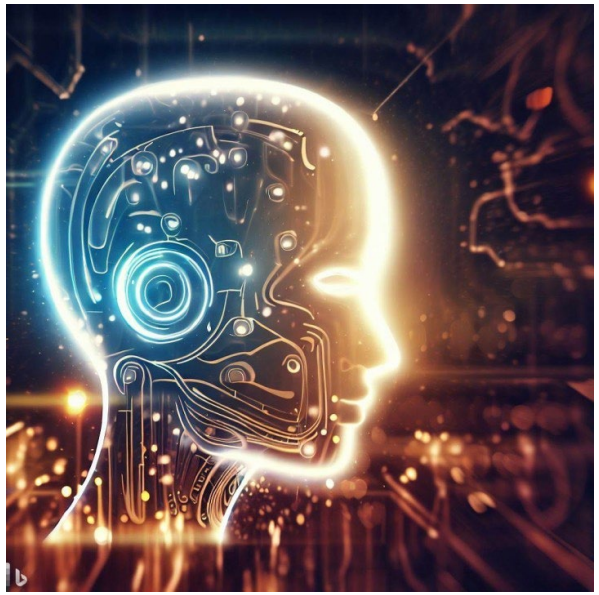
Meget svært/umuligt at formulere moralske regler der er

1. Vejledende i konkrete situationer
2. Undtagelsesløse
3. Uden behov for menneskelig fortolkning

Alt hvad jeg rører skal blive til guld!

…most goals that are easy to specify will not capture the context-specific complexities of human objectives in the real world (V&H 2021, 734)

DEN KOMPETENTE MORALSKE AGENT



Kan

1. **Fortolke** moralske regler rimeligt ift. konteksten
2. **Prioritere** moralske regler rimeligt ift. konteksten
3. **Tilsidesætte** moralske regler rimeligt ift. konteksten

Giv et eksempel på en moralsk (tommelfinger)regel, der bør anvendes meget forskelligt i forskellige kontekster!

Nobody has responded yet.

Hang tight! Responses are coming in.

SKÆRPELSE AF MIDAS-PROBLEMET: VÆRDIERS SKRØBELIGHED (*VALUE FRAGILITY*)

…if an AI system gets our values even slightly wrong, it could lead to disastrous outcomes. Hence, the more we rely on powerful autonomous systems, the more important it will be for us to specify their goals with great care, ensuring that we express our objectives correctly and completely …(…)…AI systems frequently find ways to maximize their reward functions with unintended behaviours—what Bostrom (2014: 120–124) calls 'perverse instantiations' (V&H 2021, 734)

EKSEMPEL På "PERVERS INSTANTIERING" (ibid.)

Robotstøvsugerens ordre: Sug så meget støv op fra gulvet som muligt!

Robottens løsning: Tømmer og fylder sin beholder på samme sted igen og igen

Hvad går især galt i robottens eksekvering af ordren "sug så meget støv op fra gulvet som muligt!"?

Den er ude af stand til at fortolke reglens indhold ift. dens arbejdsopgave

0%

Den er ude af stand til at underordne målet ift. vigtigere mål

0%

Den er ude af stand til at tilsidesætte ordren, når den bør

0%

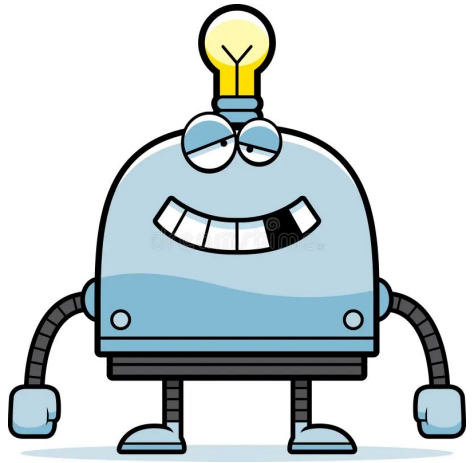
Ingen af delene

0%

LOOSEMORES ”FALLACY OF DUMB AI” - EN UDFORDRING AF ORTOGONALITETSPRINCIPPET?

If the poor machine could not understand the difference between "maximize human pleasure" and "put all humans on an intravenous dopamine drip" then it would also not understand most of the other subtle aspects of the universe, including but not limited to facts/questions like: "If I put a million amps of current through my logic circuits, I will fry myself to a crisp", or "Which end of this Kill-O-Zap Definit-Destruct Megablaster is the end that I'm supposed to point at the other guy?". Dumb AIs, in other words, are not an existential threat…(…)… my argument is that when a computer is as dumb as that, it cannot get to be as powerful as that (2012, [2])

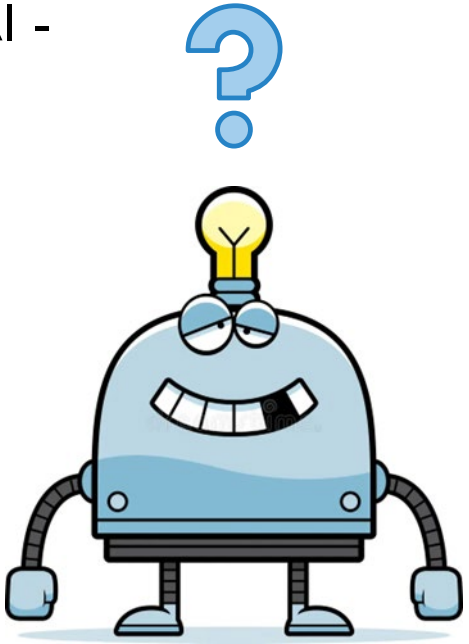
1. Hvis en AI ikke kan fortolke etiske målsætninger korrekt, så kan den heller ikke fortolke destruktive målsætninger korrekt
2. Hvis en AI ikke kan fortolke destruktive målsætninger korrekt, så kan den ikke effektivt gennemføre destruktive planer
3. Hvis en AI ikke effektivt kan gennemføre destruktive planer, så udgør den ikke en eksistentiel trussel
4. ERGO: En “dum” AI som ikke kan fortolke etiske målsætninger korrekt, udgør ikke en eksistentiel trussel (1,2,3, Hypotetisk Syllogisme)



RÆKKEVIDDEN AF LOOSEMORES ARGUMENT

Loosemore adresserer hverken **systemiske risici** eller **misbrugsrisici** relateret til AI - kun **accidentielle risici** relateret til overgreb på menneskeheden og dens ressourcegrundlag

Og tilsyneladende kun accidentielle risici relateret til **fejlfortolkning** af etiske principper, ikke til **rangordning** eller **tilsidesættelse** af principper



En meget barmhjertig læsning af Loosemores konklusion altså:

En AI der er for "dum" til at fortolke etiske principper korrekt, vil også være for "dum" til at gennemføre en plan om destruktionen af menneskeheden og/eller dens nødvendige ressourcer

INDVENDINGER MOD LOOSEMORE – 1. NATURALISTISK FEJLSLUTNING?

…Hume's (1739) longstanding is–ought problem lends support for the idea [of Orthogonality]: if one cannot infer normative statements from descriptive ones, then however intelligent a system is, it may never arrive at any moral facts

- V&H 2021, 733

M.a.o: *Pace* Loosemore kan en superintelligent robot være komplet moralsk uvidende!

SVAR:

En AI vil under alle omstændigheder have **naturaliserede** "værdier" – dermed intet "fact-value gap" !

Den vil ikke på menneskelig vis **forstå** at menneskelig velfærd er **godt**

MEN **simulere** en moralsk agent, ved at **maksimere** menneskelig velfærd

INDVENDINGER MOD LOOSEMORE – 2. DISANALOGI?

Loosemore behøver præmissen:

1. Hvis en AI ikke kan fortolke etiske målsætninger korrekt, så kan den heller ikke fortolke destruktive målsætninger korrekt

Men dette forudsætter at fortolknings-opgaverne er **sammenlignelige** i sværhedsgrad!

PROBLEM:

Er begreber som VELFÆRD, LYKKE, LIVSKVALITET osv. ikke væsentligt sværere at fortolke og operationalisere end fx ELEKTRISK BATTERI og VARME?!

NÆSTE GANG

Maskinetik:

Hvordan bygger vi etiske robotter?