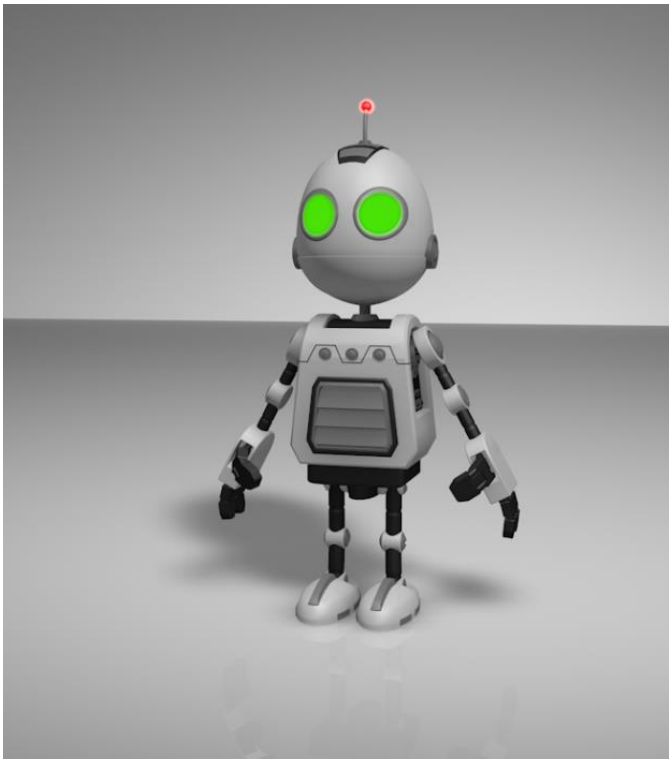
Marc Worsøe-Petersen
Studienummer: 20183482

Kunstig intelligens og moralsk status



Aalborg Universitet, Anvendt Filosofi

Kandidatspeciale, 25/10 - 2023

Vejleder: Henrik Jøker Bjerre

Antal anslag: 120.118

Marc Worsøe-Petersen
Studienummer: 20183482

Abstract

In recent years, artificially intelligent systems have increasingly become a part of our everyday lives. These systems are usually seen as tools or instruments that help us complete various tasks. However, some philosophers have begun to examine more seriously the idea of AI systems as moral subjects. This usually hinges on the assumption that future AI systems will be capable of possessing certain properties that are deemed morally important.

This master thesis explores the idea of AI systems as moral subjects by posing the following question: What criteria must an artificially intelligent system meet in order to have moral status? To answer this question, I draw on the works of several philosophers dealing with questions regarding moral status, and more specifically, the moral status of AI systems. In addition to this, I draw on the works of computer scientists Erik Learned-Miller and Miroslav Kubat.

After an initial overview of artificial intelligence and the concept of moral status, the thesis will present some of the proposed criteria an artificially intelligent system must meet in order to have moral status. These criteria include sentience, consciousness, autonomy, or the ability to set goals, advanced cognitive capabilities, (social) relations, and behavior.

Following this presentation, I will critically discuss and compare the main arguments for each position. This critical discussion will show that the arguments for sentience as the criterion an artificially intelligent system must meet in order to have moral status are the most convincing. What matters is whether or not the AI system in question is capable of experiencing emotional or physical pleasure or pain. An artificially intelligent system must have sentience in order to have any interests. Positions that propose non-sentient interests or goals as sufficient for moral status rely on a notion of interests and goals that seems nonsensical and counterintuitive.

It is then argued that the possession of advanced cognitive capabilities is an underinclusive criterion for moral status. It is, however, a reasonable criterion for moral agency.

Following this, it is shown that the relational approach to questions regarding the moral status of AI systems ultimately results in moral relativism.

Lastly, ethical behaviorism is shown to be overinclusive due to the consequences of adopting its proposed comparative principle.

Indholdsfortegnelse:

1. Indledning og problemformulering

Kunstig intelligens - også kaldet AI (Artificial Intelligence) - udvikler sig med hastige skridt og har dybtgående indvirkninger på vores hverdag og vores samfund. Kunstig intelligens bliver mere og mere allestedsnærværende, og vi bruger kunstig intelligens i mange aspekter af vores hverdagsliv. Vi bruger stemmestyrede assistenter som Siri og Alexa. Når vi bruger GPS på vores smartphone, findes den hurtigste rute via kunstig intelligens. Senest er chatværktøjet ChatGPT kommet til, som yderligere har skubbet til vores stigende brug af kunstig intelligens (Faktalink, 2023).

Også fra politisk side ses store muligheder i kunstig intelligens. Finans- og Erhvervsministeriet udgav i foråret 2019 en national strategi for udviklingen af kunstig intelligens i Danmark med en fælles vision om, at "Danmark skal gå forrest med ansvarlig udvikling og anvendelse af AI".
Senere samme år indgik Regeringen, Kommunernes Landsforening og Danske Regioner en aftale om at støtte kunstig intelligens med 200 mio. kr. i perioden 2019-2022 og til en række udvalgte projekter, hvor kunstig intelligens skal afprøves inden for sundhedsområdet, social- og beskæftigelsesområdet samt til tværgående sagsbehandling (Lindskov, Ibrahim, Thomsen, Bell & Kruse, 2020, s. 2-3).

Den offentlige diskurs omkring kunstig intelligens tager typisk udgangspunkt i en forståelse af kunstig intelligens som et værktøj, der kan bruges instrumentelt til at effektivisere og forbedre udførelsen af bestemte opgaver. Men er det muligt, at kunstig intelligens kan være mere end blot et værktøj til udførelse af menneskelige mål og interesser?
I den filosofiske litteratur er der et stigende antal af filosoffer, som tager dette spørgsmål seriøst (DeGrazia, 2022; Gibert & Martin 2022; Ladak, 2023; Neely, 2014; Mosakas, 2021; Coeckelbergh, 2014; Danaher, 2020, m.m.). Kan et kunstigt intelligent system være en moralsk agent? Eller kan det have moralsk status? Og hvad skal der til, for at dette kan blive en realitet? Det er bl.a. spørgsmål som disse, der beskæftiger filosoffer inden for området. Disse spørgsmål stilles dog oftest med det forbehold, at det højst sandsynligt først er i fremtiden, at AI-systemer er i stand til fx at opnå moralsk status.

Kunstig intelligens er altså et område med både nuværende/aktuel relevans og fremtidig relevans, og det er på den baggrund, at jeg er nået frem til følgende problemformulering:

Hvilke kriterier skal et kunstigt intelligent system opfylde for at have moralsk status?

Specialets struktur

Specialet er struktureret således, at det første hovedafsnit vil give et kort overblik over, hvad kunstig intelligens er. Afsnittet fremstår en anelse teknisk, men jeg mener, at det er nødvendigt at præsentere centrale områder inden for kunstig intelligens, for at kunne bibringe en forståelse af, hvad et kunstigt intelligent system er.
Dernæst vil der være en præcisering af begrebet "moralsk status", eftersom netop dette begreb er af væsentlig betydning for specialet.
Efterfølgende vil jeg redegøre for nogle af de kriterier, som er blevet fremsat som værende nødvendige og/eller tilstrækkelige betingelser for, at et kunstigt intelligent system kan have moralsk status. I redegørelsen vil jeg inddrage adskillige teoretikere, der behandler spørgsmål om moralsk status og mere specifikt spørgsmål om kunstigt intelligente systemers moralske status. Herefter vil jeg, med udgangspunkt i problemformuleringen, forholde mig kritisk til de anvendte teoretikeres positioner ved at diskutere, modstille og vurdere deres respektive kernepointer og argumenter.
Til sidst vil jeg i min konklusion sammenfatte redegørelsen og diskussionen og besvare min problemformulering.

Kort begrebsafklaring

Det engelske begreb "sentience" spiller en stor rolle i specialet og dækker over en entitets evne til at føle/opleve noget som positivt/fornøjeligt eller negativt/lidelsesfyldt. Jeg har ikke været i stand til at finde et dansk begreb, som dækker over de relevante elementer. Derfor vil jeg i specialet bruge det engelske begreb. Dette betyder også, at jeg undervejs i specialet bøjer ordet, således at det fremstår mere meningsfuldt ved siden af danske ord. Jeg vil fx bruge ord som sentient, sentiente, ikke-sentient, ikke-sentiente.

2. Hvad er kunstig intelligens?

Eftersom dette speciale beskæftiger sig med etiske spørgsmål vedrørende kunstig intelligens, er det naturligvis hensigtsmæssigt at danne sig et overblik over, hvad begrebet kunstig intelligens egentlig dækker over – så hvad er kunstig intelligens?

Definitioner af kunstig intelligens

Der er ikke en fast og generelt accepteret definition af kunstig intelligens. De fleste definitioner vil imidlertid have nogle fælles kendetegn. Matematikeren og datalogen John McCarthy beskrev, i 1955, kunstig intelligens på følgende måde: ”The artificial intelligence problem is taken to be that of making a machine behave in ways that would be called intelligent if a human were so behaving.” (McCarthy, Minsky, Rochester & Shannon, 1955). En anden definition kommer fra Andreas Kaplan og Michael Haenlein, der definerer kunstig intelligens som ”A system’s ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation” (Kaplan & Haenlein, 2019, s. 15). En lignende definition fremsættes af en ekspertgruppe nedsat af Europa-kommissionen: “Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.” (European Commission, 2018, s. 1). Ydermere påpeges det, at disse AI-systemer kan opdeles i to varianter: Systemer som er rent software-baserede, og systemer som er indlejret i hardwareenheder (European Commission, 2018, s. 1).

De to sidstnævnte definitioner understreger særligt evnen til at aflæse/analysere sine omgivelser, og derudfra handle eller målsætte. Der er altså tale om en vis fleksibilitet og evne til at aflæse og tilpasse sig sine omgivelser/eksternt data. Formår et system at handle/målsætte ud fra denne information, er der tale om et kunstigt intelligent system.

Førstnævnte definition er ikke lige så specifik, hvad angår et AI-systems evner. Her fremhæves en generel evne til at opføre sig/handle på måder, som vi ville kategorisere som intelligente, hvis et menneske opførte sig/handlede på disse måder. På trods af denne definitionsforskel kunne man påpege, at evnen til at aflæse/analysere sine omgivelser og derudfra handle eller målsætte netop ville kategoriseres som intelligent, hvis et menneske udviste denne evne.

En bred og mere overordnet definition kunne være Vincent C. Müllers definition: ”The notion of “artificial intelligence” (AI) is understood broadly as any kind of artificial computational system that shows intelligent behaviour, i.e., complex behaviour that is conducive to reaching goals.” (Müller, 2020). Kunstig intelligens vedrører altså maskiner/computersystemer, som udviser intelligent opførsel/adfærd – typisk forstået som kompleks opførsel/adfærd der er befordrende for indfrielsen af mål.

Turing-testen

I 1950 præsenterede den britiske matematiker Alan Turing en test, hvis formål var at teste en maskines evne til at udvise intelligens på et menneskeligt niveau. Turing kaldte testen for et imitationsspil. Testen/legen involverer oprindeligt tre personer: en mand (A), en kvinde (B) og en kvindelig eller mandlig dommer/udspørger (C).

Dommeren/udspørgeren befinder sig i et adskilt rum. Dommerens/udspørgerens opgave er at stille spørgsmål til henholdsvis manden og kvinden (udelukkende gennem skriftlig kommunikation) og ud fra deres respons, gætte hvem der er manden, og hvem der er kvinden. A's opgave er få dommeren/udspørgeren til at gætte forkert. B's opgave er modsat at hjælpe dommeren/udspørgeren.

Hvis (A) udskiftes med en maskine, bliver dommerens/udspørgerens opgave en helt anden: han/hun skal gætte hvem der er maskinen og hvem der er mennesket. Spørgsmålet er så, om dommeren/udspørgeren gætter forkert lige så ofte, når (A) er en maskine, som når (A) er et menneske.

Er dette tilfældet, konkluderes det, at maskinen udviser intelligens på et menneskeligt niveau. Eftersom dommeren/udspørgeren kan stille hvilket som helst spørgsmål, skal maskinen, for at overbevise dommeren/udspørgeren om, at den er et menneske, demonstrere at den har basal viden og en grundlæggende evne til at rationalisere (Hauser, u.å.).

Turing forudsagde i 1950, at det i år 2000 ville være muligt at programmere computere, som ville være så gode til imitationsspillet, at en gennemsnitlig dommer/udspørger ikke ville have mere end 70% chance for at gætte rigtigt efter fem minutters spørgsmål og respons. En forudsigelse der viste sig at være forkert. Ved Loebner Prize konkurrencen i år 2000 var maskinerne så dårlige til

imitationsspilet, at en gennemsnitlig dommer/udspørger havde en 100% chance for at gætte rigtigt efter fem minutters spørgsmål og respons (Hauser, u.å.).

Turing-testens fokus på en maskines evne til at udvise eller imitere menneskelig intelligens falder i tråd med den definition, John McCarthy giver af kunstig intelligens fem år senere. Her er det også en maskines evne til at opføre sig/handle på måder, som vi ville kategorisere som intelligente, hvis et menneske opførte sig/handlede på disse måder, der er væsentligt.

Stærk og svag kunstig intelligens

De fleste nuværende AI-systemer kan kategoriseres som svag AI.
Et svagt AI-system er programmeret til at udføre en snæver og specifik opgave. Sådan et system er egnet til at håndtere et specifikt problem. Et svagt AI-system er begrænset i den forstand, at det ikke kan løse andre problemer end det, det er programmeret til at løse (Bartneck, Lütge, Wagner & Welsh, 2021, s. 10).
Stærk AI kan også betegnes som generel kunstig intelligens og refererer til et system, der, i modsætning til et svagt AI-system, er i stand til at løse adskillige problemer, som ikke ligger inden for et begrænset problemfelt. Stærk AI/generel kunstig intelligens er på nuværende tidspunkt blot et mål, da ingen AI-systemer i dag kan betegnes som stærk AI/generel kunstig intelligens (Bartneck et al., 2021, s. 10).

Eksempler på AI-systemer og problemfelter inden for AI

Vidensrepræsentation er et vigtigt problemfelt inden for kunstig intelligens. Som navnet antyder, beskæftiger feltet sig med repræsentation af viden. Mere specifikt handler det om at repræsentere information på måder, der er ”forståelige” eller tilgængelige for et computersystem, således at computeren kan organisere og anvende denne information (Bartneck et al., 2021, s. 10). Ekspertsystemer er AI-systemer, som, på baggrund af vidensrepræsentation, er i stand til at besvare spørgsmål eller løse specifikke problemer inden for et bestemt domæne. Disse systemer kaldes ekspertsystemer, da deres formål er at repræsentere og anvende information/viden fra en menneskelig ekspert (Bartneck et al., 2021, s. 10). Computer vision er et andet felt inden for kunstig intelligens, som beskæftiger sig med at omdanne data fra et kamera til vidensrepræsentationer

(Bartneck et al., 2021, s. 10). Et AI-system med computer vision er altså i stand til at danne vidensrepræsentationer af visuelle data.

Maskinlæring er et felt inden for kunstig intelligens, som beskæftiger sig med skabelsen af algoritmer, der i henhold til en bestemt type af opgaver/problemer lærer af erfaring og bruger denne erfaring til at forbedre deres håndtering af nye problemer af samme type (Bartneck et al., 2021, s. 11). Siden starten af det 21. århundrede har mange forskere inden for kunstig intelligens primært forbundet kunstig intelligens med maskinlæring (Gordon & Nyholm, u.å.). AI er dog et bredere begreb og dækker bl.a. også over de systemer og problemfelter, der blev nævnt ovenfor.

Maskinlæring opdeles typisk i tre forskellige former for læringsalgoritmer: Supervised learning, unsupervised learning og reinforcement learning (Bartneck et al., 2021, s. 11-12). Den følgende redegørelse vil imidlertid kun beskrive de to førstnævnte læringsalgoritmer. I supervised learning er målet ofte at udvikle en klassifikationsalgoritme. Denne algoritme vil automatisk kunne klassificere en ny dataprøve. Algoritmen/maskinen skal dog først oplæres/trænes, hvilket sker gennem supervised learning (Bartneck et al., 2021, s. 11).
Maskinen præsenteres for to datasæt: et træningssæt og et testsæt. I træningssættet er dataet/inputtet på forhånd klassificeret. I testsættet præsenteres maskinen for data/input af samme klasse, men som ikke er blevet klassificeret på forhånd. Målet er, at maskinen kan udvikle en regel eller en procedure, der kan klassificere dataet/inputtet i testsættet ved at analysere det klassificerede data/input, den blev præsenteret for i træningssættet (Learned-Miller, 2014, s. 2).

I unsupervised learning præsenteres maskinen for et træningssæt, hvor dataet/inputtet ikke er klassificeret på forhånd. Dette står altså i modsætning til supervised learning, hvor dataet/inputtet i træningssættet er klassificeret på forhånd (Kubat, 2021, s. 297). I unsupervised learning er fokusset rettet mod en maskines evne til at forstå/gennemskue mønstre og relationer i dataet/inputtet (Bartneck et al., 2021, s. 11).

Robotter og kunstig intelligens

Som tidligere nævnt, kan AI-systemer opdeles i to varianter: Systemer som er rent software-baserede, og systemer som er indlejret i hardwareenheder. Et eksempel på et AI-system der er indlejret i en hardwareenhed, er en kunstigt intelligent robot (European Commision, 2018, s. 1). Det

er imidlertid vigtigt at pointere, at robotter og kunstig intelligens er to forskellige ting. AI-systemer kan være rent software-baserede. Robotter er derimod fysiske maskiner og er som udgangspunkt ikke AI-systemer. Kombinationen af de to er dog mulig.
Der er altså systemer, som udelukkende er AI, systemer, som udelukkende er robotter, og systemer, som er begge dele (Müller, 2020). Denne distinktion er væsentlig, da størstedelen af den filosofiske litteratur, som vil blive præsenteret senere i specialet, beskæftiger sig specifikt med kunstigt intelligente robotter. Formålet med specialet er derimod at beskæftige sig med AI-systemer generelt.

Efter dette overblik over kunstig intelligens er det på tide at dykke ned i nogle af de etiske/moralfilosofiske perspektiver på kunstig intelligens. Følgende hovedafsnit vil gøre netop dette.

3. AI-systemer som moralske subjekter

I indledningsafsnittet og problemformuleringen blev det tydeliggjort, at nærværende speciale omhandler kunstigt intelligente systemers moralske status. Når det drejer sig om forholdet mellem kunstig intelligens og etik/moralfilosofi, er det hensigtsmæssigt at præcisere, hvilket perspektiv der indtages i belysningen af dette forhold. Her er det formålstjenligt at tage udgangspunkt i Vincent C. Müllers opdeling mellem AI-systemer som objekter og AI-systemer som subjekter (Müller, 2020).
AI-systemer, som værktøjer (objekter), som mennesker skaber og anvender, er forbundet med adskillige etiske problemstillinger. AI-systemer, som subjekter, er forbundet med andre etiske spørgsmål og problemstillinger (Müller, 2020). Menneskers brug af AI-systemer (som objekter) involverer etiske problemstillinger, såsom maskin-bias/AI-systemer med bias og indsamling og analyse af privat data gennem AI-systemer (Müller, 2020).
Specialet vil imidlertid beskæftige sig med AI-systemer som subjekter. Denne tilgang er ofte forbundet med machine ethics (Müller, 2020). Machine ethics beskæftiger sig med at skabe maskiner, der er i stand til at følge etiske principper og som yderligere er i stand til at handle/træffe beslutninger ud fra disse etiske principper:

”The ultimate goal of machine ethics, we believe, is to create a machine that itself follows an ideal ethical principle or set of principles; that is to say, it is guided by this principle or these principles in

decisions it makes about possible courses of action it could take . . . Machine ethics is concerned with adding an ethical dimension to machines." (Anderson & Anderson, 2007, s. 15).

Målet er altså at skabe en form for moralsk agent. Det ultimative mål er at skabe en såkaldt "explicit ethical agent" (Anderson & Anderson, 2007, s. 15). Nærværende speciale vil dog primært fokusere på AI-systemer som moralske patienter. Der er væsentlige forskelle mellem en moralsk agent og en moralsk patient.
En moralsk agent forstås typisk som en rationel entitet, der er i stand til frit at ræsonnere og handle ud fra moralske principper/regler (Himma, 2009, s. 22-24). Med disse kapaciteter følger et moralsk ansvar. En moralsk agent holdes ansvarlig for sine handlinger. (Himma, 2009, s. 21).
En moralsk patient er derimod ikke moralsk ansvarlig for sine handlinger. Moralske patienter er ikke i stand til frit at ræsonnere og handle ud fra moralske principper/regler. Derfor kan de ikke holdes moralsk ansvarlige (Himma, 2009, s. 21). Dette er dog ikke ensbetydende med, at moralske patienter er uden moralsk status: "A moral patient is thus a being who possesses some moral status—i.e. is owed moral duties and obligations, and is capable of suffering moral harms and experiencing moral benefits." (Danahar, 2019, s. 132). Moralske agenter er altså moralsk ansvarlige for deres behandling af moralske patienter.

Spørgsmålet er så, hvad der skal til, for at et AI-system kan siges at være en moralsk patient. David J. Gunkel og Joanna Bryson formulerer spørgsmålet på følgende måde:

"Questions of moral patiency, by contrast, involve deciding whether and to what extent machines might be understood as recipients of moral action and thereby an entity who possesses what are often called "rights" requiring protection and/or respect." (Gunkel & Bryson, 2014, s. 7).

Nærværende speciale vil beskæftige sig med et lignende spørgsmål. Der er dog den forskel, at spørgsmålet ændres, således at det at være moralsk patient i stedet for forstås som det at have moralsk status. Dette fremgår af specialets problemformulering. Men hvad vil det overhovedet sige at have moralsk status? Netop det spørgsmål vil blive behandlet nu.

4. Hvad er moralsk status?

En entitet har moralsk status, hvis entiteten betyder noget, moralsk set, for dens egen skyld (Jaworska & Tannenbaum, 2013). Dette betyder, at der er moralske grunde til, hvordan entiteten skal behandles, og disse grunde vedrører entiteten selv (Jaworska & Tannenbaum, 2013). Dette ligner Frances Kamms definition af moralsk status:

”So, we see that within the class of entities that count in their own right, there are those entities that in their own right and for their own sake could give us reason to act. I think that it is this that people have in mind when they ordinarily attribute moral status to an entity . . . I shall say that an entity has moral status when, in its own right and for its own sake, it can give us reason to do things such as not destroy it or help it.” (Kamm, 2007, s. 229).

Kamm påpeger ligeledes, at en entitet har moralsk status, hvis der, for entitetens egen skyld, er grunde til at behandle den på en bestemt måde.

Der er grundlæggende to måder at forstå, hvad det vil sige at have moralsk status på. Inden for den utilitaristiske ramme har en entitet moralsk status, hvis dens velfærdsrelaterede interesser inkluderes i den utilitaristiske kalkule, der fastsætter, hvilken handling der producerer/skaber mest mulig nytte (Jaworska & Tannenbaum, 2013). Inden for den ikke-utilitaristiske ramme har en entitet moralsk status, hvis der, for entitetens egen skyld, er grunde til at handle. Disse grunde kan ikke tilsidesættes på baggrund af en beregning af de bedst mulige konsekvenser (Jaworska & Tannenbaum, 2013).

Der er forskellige opfattelser af, hvilke egenskaber en entitet skal besidde for at have moralsk status. Disse egenskaber inkluderer bl.a. avancerede kognitive evner som autonomi og rationalitet, bevidsthed, sentience (en entitets evne til at opleve/føle noget som positivt/fornøjeligt, eller negativt/lidelsesfyldt), m.m. (Jaworska & Tannenbaum, 2013). Disse egenskaber vil ikke blive uddybet yderligere her, da de hver især vil blive redegjort for i specialets redegørende del.
Udover disse egenskabsbaserede tilgange til moralsk status er der også mere alternative tilgange. Der er bl.a. fortalere for, at relationer (Coeckelbergh, 2014) eller adfærd/opførsel (Danaher, 2020)

er afgørende for moralsk status. Disse alternative tilgange vil også blive uddybet i specialets redegørende del.

Et oplagt eksempel på en entitet, der har moralsk status, er et menneske. Der er dog forskellige opfattelser af, hvilke egenskaber der er moralsk relevante i forhold til menneskers moralske status (sentience, avancerede kognitive evner, vores art osv.). Der er imidlertid også problematikker forbundet med bestemmelsen af de moralsk relevante egenskaber. Hvis det fx antages, at avancerede kognitive evner udgør den moralsk relevante egenskab, som begrunder menneskers moralske status, vil visse mennesker blive ekskluderet. Mennesker med mentale handicap eller spædbørn kan næppe leve op til denne nødvendige betingelse (Jaworska & Tannenbaum, 2013). En mulighed er at sænke barren og i stedet for hævde, at kapaciteten for at udvikle avancerede kognitive evner også er tilstrækkeligt (Jaworska & Tannenbaum, 2013). Der er også problemer forbundet med andre egenskaber, men disse vil ikke blive uddybet nærmere her (Jaworska & Tannenbaum, 2013).

Modsat os mennesker, er det næppe kontroversielt at påstå, at en sten ikke har moralsk status. En sten har ingen velfærdsinteresser, som kan inkluderes i den utilitaristiske kalkule. Der er heller ikke, for stenens egen skyld, nogen grunde til at handle.

En afgørende forskel i diskussioner om moralsk status er forskellen mellem entiteter, der har moralsk status, og entiteter som har moralsk betydning men ingen moralsk status (Mosakas, 2021, s. 430). En entitet kan have ekstrinsisk, instrumentel emotionel eller anden moralsk betydning. Vi har indirekte pligter over for disse entiteter. Modsat har vi direkte pligter over for entiteter med moralsk status (Mosakas, 2021, s. 430). Et klassisk eksempel er Immanuel Kants opfattelse af dyr. Modsat hvad de fleste mener i dag, argumenterede Kant for, at dyr ikke havde moralsk status. Han mente derimod, at vi havde indirekte pligter over for dyr (Gruen, 2003). Følgende er fra Kants Lectures on Ethics:

”If a man shoots his dog because the animal is no longer capable of service, he does not fail in his duty to the dog, for the dog cannot judge, but his act is inhuman and damages in himself that humanity which it is his duty to show towards mankind. If he is not to stifle his human feelings, he

must practice kindness towards animals, for he who is cruel to animals becomes hard also in his dealings with men." (citat i Gruen, 2003).

Ifølge Kant er det forkert, at manden skyder sin hund. Dette er dog ikke, fordi hunden har moralsk status. Det er forkert, fordi han har en indirekte pligt over for hunden. Det, han i virkeligheden skader, er sig selv – hans egen menneskelighed. Vi skal udvise godhed over for dyr, fordi det modsatte – dårlig behandling af dyr – kan påvirke vores behandling af andre mennesker. I den forstand har vi ikke en direkte pligt over for dyr, men kun en indirekte pligt.

Efter dette overblik over begrebet "moralsk status" er det på tide at gå i dybden med nogle af de kriterier, der er blevet fremsat som værende nødvendige og/eller tilstrækkelige betingelser for, at et kunstigt intelligent system kan have moralsk status.

5. Redegørelse – kriterier for moralsk status

Nærværende hovedafsnit udgør specialets redegørende del. Afsnittet har til formål at redegøre for forskellige filosofiske teorier og tilgange til spørgsmålet om AI-systemers nutidige og fremtidige moralske status. Redegørelsen vil gennemgå nogle af de kriterier, som er fremtrædende i diskussioner om moralsk status generelt. I den udvalgte litteratur, som specialet ønsker at redegøre for, undersøges/behandles disse kriterier i forhold til kunstigt intelligente systemer.

De tre første afsnit vil beskrive egenskabsbaserede tilgange til spørgsmål om moralsk status. Denne tilgang er den mest udbredte inden for diskussioner om moralsk status og består grundlæggende i, at det at have moralsk status er et spørgsmål om at besidde de moralsk relevante egenskaber. De to sidste afsnit vil beskrive to alternative tilgange til spørgsmål om moralsk status.

5.1 Sentience som kriterie for moralsk status

Nærværende afsnit har til formål at redegøre for en tilgang til spørgsmålet om kunstigt intelligente systemers moralske status, der anser sentience som en nødvendig og tilstrækkelig betingelse for moralsk status. Fortalere for denne tilgang argumenterer for, at det centrale spørgsmål i

diskussionen om AI-systemers moralske status er følgende: Er AI-systemet i stand til at opleve/føle noget som negativt/lidelsesfyldt eller positivt/fornøjeligt? Eller med andre ord: Har AI-systemet sentience?
Redegørelsen vil tage udgangspunkt i en artikel af den amerikanske filosof David DeGrazia.

David DeGrazias interessebaserede tilgang

I artiklen Robots with moral status? (2022) behandler den amerikanske filosof David DeGrazia spørgsmålet om kunstig intelligens' moralske status. Mere specifikt behandler DeGrazia spørgsmålet om AI-robotters moralske status. Argumenterne og overvejelserne i artiklen gør sig imidlertid også gældende for andre kunstigt intelligente systemer. Dette pointerer DeGrazia også (DeGrazia, 2022, s. 73-74). DeGrazia definerer kort robotter og kunstig intelligens på følgende måde:

”By robots I mean programmable machines that interact with their environment using sensors and that can perform actions at least somewhat independently of their programmers. AI involves the development of computer programs that can perform tasks that would otherwise require human (or at least organic) intelligence.” (DeGrazia, 2022, s. 74).

Ovenstående definition af kunstig intelligens læner sig tæt opad, John McCarthys tidligere nævnte definition. Et computerprograms evne til at klare opgaver, som ellers ville kræve menneskelig intelligens, ligner McCarthys beskrivelse af kunstigt intelligente maskiner, som maskiner der opfører sig/handler på måder, vi ville kalde intelligente, hvis et menneske opførte sig/handlede på disse måder.

DeGrazia antager at kunstig intelligens og robotter vil være langt mere avancerede i fremtiden. Deres nuværende evner og kapaciteter overgår, hvad vi var i stand til at forestille os for en generation siden. Intet tyder på, at denne teknologiske udvikling har nået grænsen for, hvad vi er i stand til at fremstille/producere. Dette åbner op for spørgsmålet om, hvad der skal til, før vi kan begynde at tale om kunstigt intelligente robotter eller andre AI-systemer, som entiteter med moralsk status. Det er med andre ord et spørgsmål om, hvilke egenskaber eller evner kunstigt intelligente systemer skal besidde, for at de kan have moralsk status (DeGrazia, 2022, s. 73-74).

Sentience og interesser som de afgørende kriterier for moralsk status

Det afgørende/bestemmende kriterie, som adskiller entiteter med moralsk status fra entiteter uden moralsk status, er, ifølge DeGrazia, sentience. Det, som er afgørende, er, med andre ord, hvorvidt en entitet er i stand til at opleve/føle noget som negativt/lidelsesfyldt eller positivt/fornøjeligt (DeGrazia, 2022, s. 77). DeGrazia opdeler denne kapacitet i henholdsvis en sanselig/følende form for sentience og en rent emotionel form for sentience:

”There are different ways in which robots might be sentient. They will have sensation-based sentience if, say, they have a tactile sense and can experience mechanical, thermal, or chemical changes with a positive or negative feel—permitting pain, discomfort, or tactile pleasure . . . Robot sentience is also possible, in principle, in a purely emotional form. Imagine a robot that lacked sensory feelings but cared about accomplishing certain aims . . . Caring about achieving its aims would entail sentience, because the caring would typically generate satisfaction or frustration at the achievement or thwarting of aims.” (DeGrazia, 2022, s. 78-79).

Ovenstående citat nævner specifikt robotters sentience, men man må formode, at det samme gør sig gældende i forhold til andre entiteter med sentience (fx andre AI-systemer).

Sentience er, ifølge DeGrazia, en forudsætning for, at en entitet kan have interesser. Det er udelukkende de entiteter, der er i stand til at opleve noget som negativt/lidelsesfyldt eller positivt/fornøjeligt, som har egentlige interesser – fx en grundlæggende interesse i ikke at blive påført smerte og skade. En interesse i ikke at blive påført smerte eller skade forudsætter, at kapaciteten for disse negative/lidelsesfyldte oplevelser/følelser er til stede. Kapaciteten for positive/fornøjelige oplevelser/følelser forudsættes ligeledes; dvs. der er en velfærd eller et velvære, som entiteten har en interesse i at bevare og beskytte imod smerte eller skade (DeGrazia, 2022, s. 75-76). Entiteter som mangler sentience har ingen interesser. Det er meningsløst at tage højde for, hvorvidt fx en sten har en interesse i ikke at blive påført smerte eller skade, eller hvorvidt en bil har en interesse i at nå fra en destination til en anden. Hverken stenen eller bilen er i stand til at opleve/føle noget som negativt/lidelsesfyldt eller positivt/fornøjeligt, hvilket udelukker muligheden for, at de har disse interesser (DeGrazia, 2022, s. 77).

Det at være en entitet med moralsk status er, ifølge DeGrazia, et spørgsmål om at have interesser – og dermed sentience – som tages stilling til eller medregnes i moralske agenters behandling eller interaktion med den pågældende entitet: ”X has moral status if and only if (1) X has interests, (2) moral agents have obligations regarding their treatment of X, and (3) these obligations are responsive to X’s interests.” (DeGrazia, 2022, s. 75-76).

Det er altså for entitetens egen skyld, i form af dens interesser, at den medregnes/tages stilling til i moralsk henseende. Er det ikke for entitetens egen skyld, vil den blot medregnes instrumentelt, eller måske slet ikke medregnes (DeGrazia, 2022, s. 75).

Vender vi tilbage til de to førnævnte eksempler – stenen og bilen – står det klart, at de, på baggrund af deres mangel på sentience og interesser, ikke har nogen moralsk status. Moralske agenter kan ikke handle ud fra stenens eller bilens egen skyld. Det er kun muligt at handle ud fra en entitets egen skyld, hvis denne entitet besidder et perspektiv, som er moralsk relevant i kraft af dens sentience og interesser (DeGrazia, 2022, s. 76).

Det står derimod anderledes til med mennesker og dyr, som har sentience og dermed interesser. Her er der åbenlyst tale om entiteter med moralsk status (DeGrazia, 2022, s. 75-76). Spørgsmålet er så, hvorvidt kunstigt intelligente robotter har moralsk status. Det er, ligesom med alle andre entiteter, et spørgsmål om, hvorvidt kunstigt intelligente robotter har sentience og dermed interesser (DeGrazia, 2022, s. 78). Der er intet, der tyder på, at nuværende kunstig intelligens har sentience. Det er imidlertid ikke udelukket, at fremtidig kunstig intelligens kan have sentience (Degrazia, 2022, s. 80-81).

DeGrazias opdeling mellem biologisk liv og sentience

Hvorvidt et kunstigt intelligent system har moralsk status, er altså et spørgsmål om, hvorvidt dette system har sentience. Der er imidlertid en potentiel fare for, at AI-systemer, uanset om de har sentience eller ej, ikke opfattes som havende moralsk status. Denne fare udspringer af, hvad DeGrazia kalder ”biologism” (DeGrazia, 2022, s. 85). DeGrazia beskriver dette begreb, som en ny form for speciesisme (DeGrazia, 2022, s. 85). Begrebet speciesisme dækker kort sagt over en favorisering eller større moralsk hensyntagen til de interesser, medlemmer af en bestemt art har,

udelukkende på baggrund af at de er medlemmer af denne art (Singer, 1993, s. 58-59). Der er imidlertid den forskel, at ”biologism” ikke vedrører forskellen mellem arters moralske status, men forskellen mellem levende og ikke-levende entiteters moralske status. DeGrazia beskriver denne nye form for speciesisme på følgende måde:

”The new battleground will feature a new type of speciesist: one who denies that artificial entities, regardless of their capacities or other properties, can have moral status or rights. Some might argue that, because robots are not alive, they will have less moral status than living things who are otherwise relevantly similar to robots.” (DeGrazia, 2022, s. 85).

”Biologism” skal altså forstås som ethvert synspunkt, der benægter eller reducerer en entitets moralske status på baggrund af, at den pågældende entitet er ikke-levende.

Hvorvidt en given entitet er i live i biologisk forstand, er ifølge DeGrazia, ikke i sig selv bestemmende for, om denne entitet har moralsk status eller ej. Det er, hvorvidt entiteten har sentience eller ej, som er afgørende (DeGrazia, 2022, s. 76).
For at understrege denne pointe påpeger DeGrazia, at fx planter, bakterier og svampe er levende i biologisk forstand, men uden at have sentience. Disse har biologiske behov, som fx sikrer overlevelse og reproduktion. Men siden de ikke har sentience, er det ikke muligt for dem at forholde sig eller tage stilling til, hvorvidt de tilfredsstiller disse biologiske behov. Der er ingen bevidsthed, og dermed ingen positive eller negative følelser/oplevelser forbundet med disse biologiske behov. Derfor har disse organismer ingen interesser. Vi kan, som moralske agenter, ikke handle ud fra disse organismers egen skyld (DeGrazia, 2022, s. 76).

DeGrazias pointe om at det er sentience, og ikke hvorvidt noget er levende eller ikke-levende, som er relevant i forhold til moralsk status, klargøres yderligere i hans tilslutning til Nick Boström og Eliezer Yudowskys ”Principle of Substrate Non-Discrimination” (DeGrazia, 2022, s. 79; Boström & Yudowsky, 2011, s. 8). Princippet lyder som følgende: ”If two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation, then they have the same moral status.” (Boström & Yudowsky, 2011, s. 8).

Princippet understreger, at det er moralsk irrelevant, hvilket stof to forskellige entiteter er lavet af/består af. Det, der er afgørende, er, hvorvidt de har samme funktionalitet og samme bevidste oplevelse (Boström & Yudowsky, 2011, s. 8).

DeGrazias opdeling mellem bevidsthed og sentience

En væsentlig pointe er, at DeGrazia ikke blot anser sentience som en tilstrækkelig betingelse, men også som en nødvendig betingelse, for at en entitet kan have moralsk status (DeGrazia, 2022, s. 77). Denne pointe viser sig især at være interessant i et fremtidsscenarie, hvor det er muligt at forestille sig en kunstigt intelligent robot, som er bevidst, men uden nogen form for sentience:

”Suppose, however, we have good reason to believe a particular robot is conscious but not sentient. The machine apparently thinks, consciously, and has aims of a sort but cannot desire anything (in a sense of “desire” that implies caring about the object of desire), cannot experience any sensory inputs as pleasant or unpleasant, and cannot have any moods or emotional states.” (Degrazia, 2022, s. 81).

Hvis sentience er en nødvendig betingelse for, at en entitet kan have moralsk status, vil en kunstig intelligent robot, som er bevidst, men uden nogen form for sentience, ikke have nogen moralsk status. En konklusion, som DeGrazia er klar over, er kontroversiel.
Han beskriver i denne sammenhæng et udfordrende modargument mod denne konklusion: ”Imagine angel-like beings who, although entirely incapable of pleasant or unpleasant experiences, nevertheless had a sort of preference to act in accordance with the moral law or in accordance with their moral (or other) values.” (DeGrazia, 2022, s. 82). Hvis det er muligt for en bevidst entitet at have præferencer eller interesserer på trods af dens mangel på sentience, er det ligeledes muligt (og berettiget) at denne entitet har moralsk status. Det er med andre ord et spørgsmål om, hvorvidt en bevidst, men ikke-sentient entitet, er i stand til at forme præferencer eller interesser. I dette tilfælde ville disse præferencer/interesser være baseret på fx moralske værdier frem for sentience (DeGrazia, 2022, s. 81-82).
DeGrazia benægter imidlertid, at denne entitet har egentlige interesser. Hvis denne entitet har visse mål/præferencer den ønsker at efterfølge, giver det ingen mening, hvis den samtidig er fuldstændig indifferent (fordi den mangler sentience) overfor, hvorvidt det rent faktisk lykkedes den at indfri

disse mål/præferencer. Sådan en entitet vil være usårlig/upåvirkelig i den forstand, at den ikke er i stand til at opleve/føle noget som negativt/lidelsesfyldt eller positivt/fornøjeligt. Denne usårlighed står i direkte modsætning til den sårbarhed, som er forbundet med entiteter med sentience. Disse entiteter har, på baggrund af deres sentience, interesser, som kan påvirke entiteten positivt, hvis de indfries/respekteres, eller negativt, hvis de ikke indfries/ikke respekteres (DeGrazia, 2022, s. 82). Det er med andre ord ikke muligt for moralske agenter, at handle ud fra en entitets egen skyld (dvs. ud fra entitetens interesser), hvis den er fuldstændig indifferent (og derfor mangler egentlige interesser) overfor, hvorvidt den indfrier sine mål/præferencer. Derfor mangler denne bevidste, men ikke-sentient entitetet moralsk status (DeGrazia, 2022, s. 75-76).

DeGrazia benægter imidlertid ikke, at bevidsthed er af væsentlig betydning, når det drejer sig om moralsk status. Bevidsthed er en nødvendig, men ikke tilstrækkelig betingelse for at en entitet kan have moralsk status. Begrundelsen for denne nødvendighed ligger i, at bevidsthed er en nødvendig betingelse for besiddelsen af sentience. Han laver altså en skildring mellem på den ene side bevidsthed og på den anden side sentience, som er: ”The capacity for consciousness that features pleasant or unpleasant experiences.” (DeGrazia, 2022, s. 74).

Man kan stille det op på følgende måde: Bevidsthed er en nødvendig betingelse for sentience, som er en nødvendig og tilstrækkelig betingelse for moralsk status. Der er, ifølge DeGrazia, imidlertid endnu et lag af moralsk beskyttelse, som en entitet kan have. Hvad dette helt præcist indebærer, vil de to følgende underafsnit præsentere.

Kunstigt intelligente robotter med rettigheder

Efter ovenstående redegørelse er der ingen tvivl om, at sentience og det at en entitetet har interesser, udgør fundamentet for, hvad det vil sige at have moralsk status. DeGrazia bevæger sig imidlertid kort ind på et yderligere spørgsmål om rettigheder. DeGrazia nævner ikke eksplicit, at der er tale om moralske rettigheder, men man må formode, at det er moralske rettigheder, han henviser til (frem for fx juridiske rettigheder), når han bruger ordet ”rettigheder”.

Det spørgsmål som DeGrazia åbner op for er, hvorvidt kunstigt intelligente robotter skal have rettigheder. Et bekræftende svar på dette spørgsmål forudsætter, ifølge DeGrazia, at AI-robotter

har en narrativ selvbevidsthed. Dette dækker over en entitets evne til at forstå/betragte sin eksistens som en slags fortælling eller et narrativ, som er opdelt i distinkte dele. Beviser for denne evne kunne fx være spontan selvrefererende tale, påstande omhandlende AI-robottens fortid og fremtid, eller spørgsmål omkring dens eksistens og plads i universet. Besidder en kunstigt intelligent robot denne evne, skal den tildeles moralske rettigheder. DeGrazia nævner, som eksempel på disse rettigheders beskyttelse, at en forventning om maksimering af nytte ikke kan retfærdiggøre at fratage en AI-robots eksistens eller på andre måder påføre den fundamental skade (DeGrazia, 2022, s. 82-83).

Det interessante ved DeGrazias beskrivelse af AI-robotter med moralske rettigheder er, at han fremhæver yderligere evner/kapaciteter end blot evnen til at opleve/føle noget som positivt/fornøjeligt eller negativt/lidelsesfyldt. Mere specifikt fremhæver han selvbevidsthed, som den skelsættende evne/kapacitet mellem henholdsvis entiteter med moralske rettigheder og entiteter uden moralske rettigheder (DeGrazia, 2022, s. 82-83).

Kunstigt intelligente robotter med autonomi

Udover AI-robotter med moralske rettigheder, beskriver DeGrazia også muligheden for AI-robotter med autonomi. Ligesom med spørgsmålet om AI-robotters moralske rettigheder, sætter også spørgsmålet om AI-robotter med autonomi fokus på yderligere evner/kapaciteter end blot evnen til at opleve/føle noget som positivt/fornøjeligt eller negativt/lidelsesfyldt (DeGrazia, 2022, s. 83). DeGrazia beskriver de krav, en entitet må opfylde, for at der er tale om en autonom entitet, på følgende måde:

”According to the conception of autonomy I favor, an agent can act autonomously if and only if she can act (1) intentionally, (2) with sufficient understanding, (3) sufficiently freely of controlling influences, and (4) in light of her own values.” (DeGrazia, 2022, s. 83).

I lyset af disse krav påpeger han yderligere: ”My speculation is that a robot that affords us good reason to believe it is conscious, sentient, and narratively self-aware might very well meet these conditions.” (DeGrazia, 2022, s. 83).

En kunstigt intelligent robot kan altså siges at være autonom, hvis den opfylder disse krav. Men betyder det så, at en kunstigt intelligent robot, som er autonom eller som har rettigheder, har en højere moralsk status end kunstigt intelligente robotter, som blot har sentience?
Næste underafsnit vil tage udgangspunkt i netop dette spørgsmål.

Grader af moralsk status

På baggrund af DeGrazias beskrivelse af muligheden for kunstigt intelligente robotter med rettigheder og autonomi samt det faktum, at besiddelsen af disse kræver yderligere evner og kapaciteter end blot sentience, kan det fremstå som om, at DeGrazia anser moralsk status som noget, en entitet kan have i større eller mindre grad.
Yderligere beskriver DeGrazia også rettigheder som noget, der tillægges ”on top of basic moral status” (Degrazia, 2022, s. 83). Denne grundlæggende moralske status må man formode, at en entitet har, hvis den besidder sentience. Rettighederne fungerer så som et ekstra moralsk lag, der lægges ovenpå denne grundlæggende moralske status. Spørgsmålet er så, om en entitet, herunder en kunstigt intelligent robot, med dette ekstra lag har en større grad af moralsk status.
Tager man udgangspunkt i en af DeGrazias andre artikler, Moral Status As a Matter of Degree? (2008), må svaret være ja. Artiklen tager primært udgangspunkt i relationer mellem mennesker og dyr og deres respektive moralske status. Dette er dog ikke en hindring for at udvide diskussionen til også at omhandle kunstigt intelligente systemer.
Det afgørende for DeGrazia er, hvorvidt en entitet har sentience, ikke om det er et dyr, et menneske eller et AI-system. Dette gøres bl.a. tydeligt i hans gentagne brug af udtrykket ”sentient beings” (Degrazia, 2008, s. 187). Ved at bruge dette udtryk forholder han sig åbent og neutralt over for, hvad denne ”being” måtte være, så længe den/det har sentience.

Det interessante ved artiklen er, at DeGrazia argumenterer for, hvad han kalder ”The Unequal Consideration Model of Degrees of Moral Status” (DeGrazia, 2008, s. 187). Fortalere for denne model anerkender, at alle entiteter med sentience har interesser og dermed moralsk status. Visse entiteter har imidlertid en højere moralsk status, fordi deres interesser er af større moralsk betydning (DeGrazia, 2008, s. 187-188). DeGrazia betegner entiteterne, som har interesser af større moralsk betydning, som personer eller moralske agenter, og adskiller dem fra ikke-personer (DeGrazia, 2008, s. 193). Personer/moralske agenter er ifølge denne model berettiget til større

moralsk hensyntagen end ikke-personer, da deres interesser er af større moralsk betydning (DeGrazia, 2008, s. 187-188). DeGrazias begrundelse for, at personer/moralske agenter har interesser af større moralsk betydning og dermed en højere moralsk status, vil imidlertid ikke blive beskrevet yderligere her. DeGrazia synes nemlig at have ændret holdning siden da.

I bogen, A Theory of Bioethics (2021), forsvarer DeGrazia og Joseph Millum nemlig, hvad de kalder ”equal consequentialist consideration” eller ”qualified equal consideration”. Ud fra denne model har alle entiteter med sentience interesser og dermed moralsk status. Ydermere tillægges disse interesser samme moralske vægt, hvis der er tale om sammenlignelige interesser. En interesse i at undgå smerte er et eksempel på en sammenlignelig interesse(DeGrazia & Millum, 2021, s. 187). Et menneskes interesse i at undgå smerte vejer ikke moralsk tungere end en hunds interesse i at undgå samme niveau/grad af smerte. Der er i den forstand en lige moralsk hensyntagen til menneskets og hundens interesse i at undgå smerte, fordi der er tale om en sammenlignelig interesse (DeGrazia & Millum, 2021, s. 193).

Det er imidlertid i mange tilfælde berettiget at behandle et menneske og fx en hund forskelligt – oftest til menneskets fordel. Dette skyldes dog ikke, at de har forskellige grader af moralsk status, men fordi de har forskellige og ikke-sammenlignelige interesser (DeGrazia & Millum, 2021, s. 189). DeGrazia og Millum nævner bl.a. at et menneskes interesse i overlevelse ikke kan sammenlignes med en kats interesse i overlevelse:

”Do they [et menneske og en kat] have roughly the same thing of value at stake in remaining alive? We think not: the person may be expected to lose more from dying prematurely and so would be harmed more extensively by death. In addition, the human will likely have a much deeper set of psychological connections to her possible future and for this reason may be judged to be harmed more extensively by death.” (DeGrazia & Millum, 2021, s. 188).

Denne usammenlignelighed mellem, hvad DeGrazia & Millum kalder ”persons” og ”nonpersons” interesser, skyldes, at personer besidder en narrativ selvbevidsthed, der gør dem i stand til at reflektere over deres fortid og fremtid og på den måde begribe deres liv som en form for fortælling/narrativ med forskellige dele. Dette gør, at personer er i stand til at have langsigtede interesser. Personer har i den forstand ofte mere at tabe ved ikke at få deres interesser imødekommet eller respekteret (DeGrazia & Millum, 2021, s. 199-200). Dette berettiger ifølge

DeGrazia og Millum, at personer tildeles moralske rettigheder, som beskytter deres langsigtede interesser. Entiteter, som har sentience, men ingen selvbevidsthed – ikke-personer – tildeles derimod ikke disse rettigheder, da de ikke er i stand til at have langsigtede interesser (DeGrazia & Millum, 2021, s. 199-200). Denne forskel mellem personer og ikke-personer er imidlertid ikke udtryk for en forskel i moralsk status. Forskellen er derimod udtryk for en forskel i interesser (DeGrazia & Millum, 2021, s. 189).

Vender vi tilbage til spørgsmålet om kunstigt intelligente robotters moralske status, må man formode, at det samme gør sig gældende her. DeGrazia nævner det samme krav – en narrativ selvbevidsthed – for at en kunstigt intelligent robot kan tildeles rettigheder (DeGrazia, 2022, s. 82-83). Denne selvbevidste robot vil imidlertid ikke have en højere moralsk status end robotter uden selvbevidsthed (eller andre entiteter med sentience, men uden selvbevidsthed). Den vil derimod have langsigtede interesser, som berettiger et ekstra lag af beskyttelse i form af moralske rettigheder.

5.2 Ikke-sentiente og ikke-bevidste AI-systemer med moralsk status?

I de forrige afsnit har sentience stået som det skelsættende kriterie, en entitet (herunder en AI-robot eller et andet AI-system) skal opfylde, for at den kan have moralsk status.
Nærværende afsnit vil ikke bevæge sig helt væk fra sentience-kriteriet, men vil derimod stille spørgsmålstegn ved, hvorvidt sentience er en nødvendig betingelse for moralsk status. Er det muligt at have moralsk status, hvis man er bevidst, men ikke sentient? Og yderligere, er det muligt at have moralsk status, hvis man hverken er bevidst eller sentient?

Nick Boström og Eliezer Yudowsky

Filosoffen Nick Boström og forsker i kunstig intelligens Eliezer Yudowsky stiller bl.a. (indirekte) de ovenstående spørgsmål: Er sentience og bevidsthed nødvendige betingelser for moralsk status? Det gør de i lyset af muligheden for, at en fremtidig kunstig intelligens kan besidde, hvad de kalder, ”eksotiske egenskaber”. Dette udtryk dækker over muligheden for, at en fremtidig kunstig intelligens udadtil vil fremstå og opføre sig menneskelignende på trods af, at den er anderledes konstitueret. Mere specifikt nævner de muligheden for, at en fremtidig kunstig intelligens kan være

konstitueret således, at den fx besidder højt avancerede kognitive evner, men ingen sentience eller bevidsthed (Boström & Yudowsky, 2011, s. 9-10).
Ifølge Boström og Yudowsky åbner dette op for følgende spørgsmål: ”Should such a system [AI med højt avancerede kognitive evner, men uden sentience eller bevidsthed] be possible, it would raise the question whether a non-sentient person would have any moral status whatever; and if so, whether it would have the same moral status as a sentient person.” (Boström & Yudowsky, 2011, s. 10). Det er således et indirekte spørgsmål om, hvorvidt sentience er en nødvendig betingelse for moralsk status. En vigtig pointe er dertil, at der også er et yderligere spørgsmål om, hvorvidt bevidsthed er en nødvendig betingelse for moralsk status.
De følgende afsnit vil tage udgangspunkt i netop disse spørgsmål.

Bevidst kunstig intelligens med ikke-sentiente interesser og mål

Er det muligt at have interesser, mål, præferencer og ønsker, hvis ikke man besidder sentience? Ifølge filosoffen Erica L. Neely er svaret ja. I artiklen Machines and the Moral Community (2014) forsøger hun bl.a., at forsvare en interessebaseret tilgang til spørgsmålet om kunstig intelligens' moralske status. Neely retter altså, ligesom David DeGrazia, sit fokus mod en entitets interesser. Der er imidlertid den væsentlige forskel, at det, ifølge Neely, er muligt for en entitet at have interesser, selvom den mangler sentience (Neely, 2014, s. 98). Et synspunkt som DeGrazia naturligvis ville forholde sig kritisk til (DeGrazia, 2022, s. 81-82).

Sentience er, ifølge Neely, ikke en nødvendig betingelse for at en entitet kan have moralsk status. Det centrale spørgsmål er, hvorvidt en entitet har interesser. Det at være sentient repræsenterer blot en måde at have interesser på. Sentience er altså en tilstrækkelig betingelse for at have interesser. Der er imidlertid andre måder, hvorpå en entitet kan siges at have interesser og dermed moralsk status (Neely, 2014, s. 98).

For at påvise, at sentience ikke er en nødvendig betingelse for at have interesser, beskriver Neely forskellige situationer, hvor der tilsyneladende er tale om sentience-uafhængige interesser. Neelys første eksempel beskriver en person, der lider af den ekstremt sjældne sygdom, kongenital analgesi. Mennesker med denne lidelse er ikke i stand til at føle fysisk smerte. Vi skal så forestille os en situation, hvor en anden person træder på hans/hendes fod uden at have fået tilladelse til det.

Handlingen forårsager ingen fysisk smerte, eftersom personen, hvis fod trædes på, ikke er i stand til at føle fysisk smerte. Alligevel fremstår handlingen moralsk forkert. Hvis sentience udelukkende dækker over evnen til at føle/mærke fysisk smerte eller nydelse, vil man, hvis man betragter sentience som en nødvendig betingelse for moralsk status, derimod ikke anse handlingen som moralsk forkert. Hvis sentience udvides til også at omfatte emotionel smerte og nydelse, vil det derimod være muligt at fordømme handlingen, da personen med kongenital analgesi kunne opleve handlingen som emotionelt smertefuld. Om personen har sentience eller ej, er imidlertid ikke det eneste moralsk relevante spørgsmål i situationen. Hvis vi forestiller os, at personen med kongenital analgesi heller ikke er i stand til at føle/opleve emotionel smerte eller nydelse, ville det ifølge Neely stadig være forkert at træde på personens fod (Neely, 2014, s. 99).

Der synes altså at være noget, som, uafhængigt af sentience, giver personen med kongenital analgesi moralsk status. Noget, der gør det moralsk forkert at træde på personens fod uden personens samtykke, selvom personen ikke føler/oplever denne handling som fysisk eller emotionelt smertefuld. Personen kan ifølge Neely såres på måder, der ikke er betinget af sentience:

”The wrongness in our case stems from two key points. First, the action could cause damage, even if it does not cause pain. Second, since we have specified that the person does not give permission for the action, deliberately stepping on his foot violates his desire to remain unmolested . . . What is necessary for moral standing is not sentience per se but having interests; the person in our congenital analgesia example lacks sensation, but he retains interests. As it is possible to harm those interests, it is possible to harm him” (Neely, 2014, s. 99).

Personen med kongenital analgesi har altså moralsk status, fordi han/hun har interesser. Det, at træde på personens fod er moralsk forkert, fordi der ikke tages moralsk hensyn til hans/hendes interesse i at forblive uskadt. Ved at ignorere eller negligere personens interesser krænkes personens autonomi (Neely, 2014, s. 100). Netop autonomi står centralt i Neelys forståelse af moralsk status. Så centralt, at der ifølge Neely kan være entiteter, som hverken har sentience eller bevidsthed, men som udelukkende på baggrund af deres autonomi har interesser og dermed moralsk status (Neely, 2014, s. 102-103). Denne pointe uddybes yderligere i næste underafsnit.

Neely beskriver nemlig først, at bevidsthed og selvbevidsthed er tilstrækkelige betingelser for, at en entitet kan have interesser:

”Hence, while sentience certainly leads to having interests, it is not necessary for them: the joint properties of consciousness and self-awareness will also suffice. Once a being is self-aware and conscious, it is aware of its self, can desire continuation of that self, and can formulate ideas about how to live its life. It is possible to harm such a being by ignoring or thwarting those desires.” (Neely, 2014, s. 100).

Denne bevidsthed og selvbevidsthed, skal forstås som uafhængig af sentience. Dvs. sentience er ikke en nødvendig betingelse for at have interesser. Bevidsthed og selvbevidsthed er tilstrækkeligt (Neely, 2014, s. 100). Det som karakteriserer den bevidste og selvbevidste entitet er, at den ønsker sin fortsatte eksistens, og at den har præferencer, mål og ønsker om at forfølge og opfylde disse mål (Neely, 2014, s. 101). Disse kapaciteter forudsætter ikke, at entiteten har sentience. Derfor har entiteten interesser, som skal respekteres og vægtes moralsk set for entitetens egen skyld. Bevidsthed og selvbevidsthed er altså tilstrækkeligt til at have interesser og dermed moralsk status (Neely, 2014, s. 100-101). Alt dette gør sig også gældende, når det drejer sig om kunstigt intelligente systemer. Hvis et kunstigt intelligent system er bevidst og selvbevidst, har systemet mål, ønsker og interesser. Disse interesser er tilstrækkelige til, at AI-systemet har moralsk status. Moralske agenter skal derfor handle ud fra disse interesser og dermed ud fra AI-systemets egen skyld (Neely, 2014, s. 100-101).

Neely er ikke den eneste, der mener, at sentience ikke er en nødvendig betingelse for at have interesser og dermed moralsk status. Ali Ladak argumenterer ligeledes for, at bevidsthed er en tilstrækkelig betingelse for at have præferencer, mål og interesser (Ladak, ingen sidetal, 2023). Ifølge Ladak er det at have mål og præferencer begrebsmæssigt adskilt fra sentience. En entitets evne til at opleve/føle noget som positivt/fornøjeligt eller negativt/lidelsesfyldt er essentielt for at have sentience. Når mennesker og dyr opnår eller ikke opnår deres respektive mål og præferencer, er det typisk forbundet med oplevelsen eller følelsen af noget positivt eller negativt. Ifølge Ladak er denne forbindelse imidlertid ikke en nødvendig forbindelse (Ladak, 2023). Han giver et eksempel, hvor der tilsyneladende er præferencer og mål, uden at disse er forbundet med positive eller

negative følelser/oplevelser. Et skakspillende computerprogram, hvis mål er at vinde, kan siges at have en præference eller et mål, selvom positive eller negative følelser/oplevelser udebliver (Ladak, 2023). En entitet, herunder et AI-system, der er bevidst, men ikke sentient, kan have disse ikke-sentiente præferencer og mål. Derfor har disse entiteter moralsk status:

”The least controversial extension of moral standing beyond sentient entities is to conscious entities with preferences and goals, but no positive or negative feelings . . . Since they have preferences and goals, they can reasonably be described as having interests that can make things go better or worse for them, and since they are conscious, there is arguably a subject for whom those interests matter.” (Ladak, 2023).

En bevidst, men ikke sentient, entitet kan altså ifølge Ladak have interesser, der har moralsk betydning, fordi der er et subjekt, for hvem disse interesser betyder noget, selvom interesserne ikke er forbundet med positive eller negative følelser/oplevelser (Ladak, 2023).

Ifølge Neely og Ladak vil et bevidst, men ikke-sentient, AI-system altså være i stand til at have mål, præferencer, interesser og dermed moralsk status. Men hvad med et AI-system, som hverken er bevidst eller sentient? Er sådan et system i stand til at have interesser? Netop det, vil næste underafsnit omhandle.

Ikke-bevidste og ikke-sentiente AI-systemer med interesser

I forrige underafsnit blev det kort nævnt, at Neely går et skridt længere og argumenterer for, at et AI-system, der har autonomi, men ingen bevidsthed eller sentience, kan have interesser og dermed moralsk status (Neely, 2014, s. 102-103). Autonomi alene er altså tilstrækkeligt.
Neely eksemplificerer dette ved at beskrive forskellen mellem to AI-systemers målsætningsevner. Et AI-system hvis mål altid bliver bestemt af en ekstern/udefrakommende kilde, er ikke autonomt. Dette AI-system er ikke i stand til at formulere og bestræbe sine egne mål. Det vil blot efterfølge instrukser fra en ekstern kilde. Sådan et system har ingen interesser og derfor ingen krav på moralsk hensyntagen. Moralske agenter kan ikke gøre noget for systemets egen skyld (Neely, 2014, s. 102). Derimod vil et AI-system, som i visse tilfælde er i stand til at bestemme sine egne mål, besidde en grundlæggende form for autonomi. I disse tilfælde vil systemet træffe beslutninger, der ikke blot

består i, at systemet følger en algoritme eller et program. Dette AI-system har altså interesser, og derfor krav på moralsk hensyntagen. Det førstnævnte system har ingen moralsk status. Systemet, som har en grundlæggende form for autonomi, har moralsk status (Neely, 2014, s. 102-103).

Ali Ladak mener ligeledes, at "entities that are non-conscious but have sufficiently cognitively complex preferences and goals" muligvis kan have interesser, der betyder noget for entiteten. Ladak specificerer imidlertid ikke, hvad der karakteriserer kognitivt komplekse præferencer og mål (Ladak, 2023). Han påpeger blot, at der i fremtiden muligvis vil være højt kognitivt sofistikerede AI-systemer, som hverken har sentience eller bevidsthed. Ifølge Ladak har sådanne AI-systemer muligvis interesser, og et perspektiv, hvorudfra disse interesser betyder noget. Systemet har altså et perspektiv, selvom det mangler bevidsthed og sentience. Ifølge Ladak er andre mentale tilstande, såsom "beliefs", muligvis tilstrækkelige for, at en entitet har et perspektiv, hvorudfra dens interesser har betydning (Ladak, 2023).

I artiklen, Non-Human Moral Status: Problems with Phenomenal Consciousness (2023), argumenterer den amerikanske filosof Joshua Shepherd ligeledes for, at bevidsthed ikke er en nødvendig betingelse for moralsk status. I artiklen tager Shepherd udgangspunkt i ikke-menneskelig moralsk status (Shepherd, 2023, s. 148). Dette inkluderer også AI-systemer. Sherperd nævner også AI-systemer undervejs i artiklen (Shepherd, 2023, s. 150 & 153). Sheperd beskriver, at en stor del den nuværende litteratur, som beskæftiger sig med moralsk status, tilslutter sig en bevidsthedsbaseret tilgang til moralsk status. Inden for denne tilgang er der forskellige variationer (Shepherd, 2023, s. 148).
Iblandt disse variationer er der imidlertid en position, som størstedelen af filosoffer og bioetikere tilslutter sig. Denne position er karakteriseret ved, hvad Shepherd kalder "the judgment of necessity" (Shepherd, 2023, s. 150). Dette udtryk dækker over, at fortalere for denne position anser bevidsthed som en nødvendig betingelse for moralsk status (Shepherd, 2023, s. 149). Fortalere for denne position kan inddeles i tre grupper: De, der mener, at at bevidsthed er en nødvendig, men ikke tilstrækkelig betingelse for moralsk status (Shepherd, 2023, s. 148-149). De, der mener, at bevidsthed er en nødvendig og tilstrækkelig betingelse for en vis grad af moralsk status, men at der også er andre faktorer, der kan påvirke graden af moralsk status (Shepherd, 2023, s. 149). Og de, der mener, at bevidsthed er en nødvendig og tilstrækkelig betingelse for moralsk status, og at

bevidsthed er den eneste faktor, der har betydning for graden af moralsk status (Shepherd, 2023, s. 149-150). Ifølge Shepherd er bevidsthed ikke en nødvendig betingelse for moralsk status. Shepherd præsenterer tre argumenter for, at bevidsthed ikke er en nødvendig betingelse for moralsk status. Nærværende redegørelse vil fokusere på to af dem: (1) Som bevidste væsener har vi mennesker ikke adgang til ikke-bevidste tilstande. Derfor er vi ikke i stand til at vurdere, om disse tilstande indeholder noget af moralsk betydning (Shepherd, 2023, s. 151-152). (2) Der er ikke-bevidste aspekter af en entitet, som i sig selv kan være tilstrækkelige for moralsk status (Shepherd, 2023, s. 150-151).

Ifølge Shepherd er det uberettiget for os mennesker, at bestemme/vurdere, hvorvidt en ikke-bevist entitet har moralsk status (Shepherd, 2023, s. 152). Det er ikke muligt for os at tage en ikke-bevidst entitets perspektiv. Derfor er det umuligt for os at vide, om en ikke-bevidst tilstand er uden nogen form for værd/værdi for den ikke-bevidste entitets perspektiv. Vi vil altid tage udgangspunkt i vores eget perspektiv og sammenligne det med et andet (Shepherd, 2023, s. 151-152). Problemet er bare, at vores perspektiv altid er stærkt forbundet med bevidsthed:

”When we are asked whether there is any value in the mental life of a non-conscious being . . . A common method is to think about what is valuable in our own mental lives ,and to think about whether that kind of thing would be present in the mental life of a non-conscious entity . . . The problem here is that access to any value in our own mental case is very strongly correlated with consciousness.” (Shepherd, 2023, s. 151).

Siden vi mangler adgang til dette perspektiv, er det ikke berettiget, at vi, ud fra vores bevidste perspektiv, bestemmer at der ikke er noget af moralsk betydning for en ikke-bevidst entitet. Siden vi ikke kan udelukke dette, kan vi heller ikke udelukke at ikke-bevidste entiteter har moralsk status (Shepherd, 2023, s. 151-152).

Ifølge Shepherd er der visse objektive goder, som det er muligt for en ikke-bevidst entitet at stræbe efter. Shepherd nævner ”the satisfaction of desires, the acquisition of knowledge, the nurturing of relationships of care, the pursuit of long-term plans, and the realization of significant achievements” som eksempler på disse goder (Shepherd, 2023, s. 152). Tilstedeværelsen af disse mål og projekter giver entiteten en vis grad af moralsk status, da disse mål og projekter er af værdi for entitetens liv,

selvom de ikke er forbundet med nogen form for bevidsthed. Bevidsthed er altså ikke en nødvendig betingelse for moralsk status (Shepherd, 2023, s. 153). Tilstedeværelsen af disse mål og projekter i et kunstigt intelligent systems "liv" må man formode giver dette system en vis grad af moralsk status.

5.3 Avancerede kognitive evner som kriterie for moralsk status

Når det kommer til sammenhængen mellem avancerede kognitive evner og moralsk status, er det svært at komme udenom den tyske filosof Immanuel Kant. Nærværende afsnit vil derfor kort præsentere en kantiansk model eller et kantiansk perspektiv på kunstigt intelligente systemers moralske status. Dette gøres primært gennem John-Stewart Gordons og David J. Gunkels beskrivelse af en kantiansk tilgang til spørgsmålet om AI-robotters moralske status.
Nærværende speciale vil derfor begrænse sig til at gennemgå Kants opfattelse af værdighed, da dette er relevant for specialets fokus på moralsk status. Følgende underafsnit udgør derfor på ingen måde en komplet redegørelse for Kants moralfilosofi. Redegørelsen vil blot beskrive én del af Kants moralfilosofi.

Kant – rationelle væseners værdighed

I bogen, Groundwork of the Metaphysics of Morals (1785), er Kants mål at søge efter og etablere, "the supreme principle of morality" (Kant, 1998, s. 5). Dette princip – det kategoriske imperativ – formulerer Kant på følgende måde: "Act only in accordance with that maxim through which you can at the same time will that it become a universal law." (Kant, 1998, s. 31).
Som rationelle og autonome væsener er vi underlagt moralloven. Men vi er samtidigt lovgivere af selvsamme morallov. Ved at ville at en maksime bliver almen/universel lov lovgiver vi og underlægges på samme tid denne lov da den er almen/universel. Loven er almen/universel, fordi alle rationelle og autonome væsener vil at maksimen bliver almen/universel lov (Kant, 1998, s. 39-40).

Rationalitet, autonomi og dermed evnen til at handle ud fra moralloven/det kategoriske imperativ giver mennesker værdighed. Kant beskriver denne værdighed som ubetinget og usammenlignelig (Kant, 1998, s. 42-43). Rationelle væsener er i kants ord, "an end in itself" (Kant, 1998, s. 43). Dette

står i stærk modsætning til det værd væsener/entiteter, som ikke er rationelle og autonome besidder:

”Beings the existence of which rests not on our will but on nature, if they are beings without reason, still have only a relative worth, as means, and are therefore called things, whereas rational beings are called persons because their nature already marks them out as an end in itself, that is, as something that may not be used merely as a means, and hence so far limits all choice (and is an object of respect).” (Kant, 1998, s. 37).

Ikke-rationelle og ikke-autonome væsener/entiteter er ikke ”ends in themselves” (kant, 1998, s. 41). Disse væsener er midler og er kun af relativt værd. Rationelle og autonome væsener må ikke bruges udelukkende som midler, men skal altid også betragtes som et mål i sig selv (Kant, 1998, s. 37). Rationalitet og autonomi er altså nødvendige betingelser for at have ubetinget værdighed, dvs. for ikke bare at være et middel, men også altid at være et mål i sig selv.

En kantiansk tilgang til AI-robotters moralske status

Da Kant i sin tid formulerede sin moralfilosofi og mere specifikt sit syn på rationelle og autonome væsener som bærere af en ubetinget værdighed, havde han næppe kunstigt intelligente systemer i tankerne. Dette er dog ikke ensbetydende med, at Kants ideer ikke kan spille nogen rolle i nutidige diskussioner om AI-systerms nutidige og fremtidige moralske status. David J. Gunkel og John-Stewart Gordon forsøger netop at præsentere et kantiansk perspektiv på kunstigt intelligente robotters moralske status (Gordon & Gunkel, 2022).
Men er det overhovedet muligt at anvende Kants idé om rationelle og autonome væseners ubetingede værdighed på spørgsmålet om AI-robotters moralske status? Er det ikke kun mennesker, som besidder denne værdighed? Gordon og Gunkel beskriver netop denne indvending mod en kantiansk tilgang til AI-robotters moralske status:

”Kant defines the notion of dignity in terms of human dignity. Consequently, it is impossible for robots—even extremely intelligent ones—to have dignity, simply because they are not human beings. And it therefore follows that if intelligent robots do not have human dignity, then they are not entitled to moral rights.” (Gordon & Gunkel, 2022, s. 96).

Dette er imidlertid, ifølge Gordon og Gunkel, en misvisende indvending. Kants begreb om ”personhood” dækker over alle rationelle personer, og personer skal i denne sammenhæng ikke udelukkende forstås som mennesker. Derfor kan intelligente robotter, hvis de altså er rationelle, også betegnes som rationelle personer. I så fald ville disse AI-robotter besidde ubetinget værdighed (Gordon & Gunkel, 2022, s. 96).

Ifølge Kant er rationelle personer i stand til at træffe beslutninger og handle ud fra moralske fordringer. Evnen til at træffe autonome og rationelle beslutninger og dermed handle i overensstemmelse med moralske principper (moralloven/det kategoriske imperativ) giver rationelle personer ubetinget værd/værdighed (Gordon & Gunkel, 2022, s. 95). Hvis kunstigt intelligente robotter i fremtiden ligeledes er i stand til at træffe autonome og rationelle beslutninger og dermed handle i overensstemmelse med moralske principper, er de, på trods af at de ikke er mennesker, rationelle personer med ubetinget værdighed (Gordon & Gunkel, 2022, s. 98). Det er altså muligt at bestemme AI-robotters moralske status ud fra en kantiansk tilgang (Gordon & Gunkel, 20202, s. 98). Disse avancerede kognitive evner (rationalitet og autonomi) er altså, ifølge den kantianske model, nødvendige betingelser for moralsk status.

5.4 Den relationelle tilgang til spørgsmålet om moralsk status

Fælles for de mange foregående afsnit er, at spørgsmålet om kunstig intelligens' moralske status grundlæggende er et spørgsmål om, hvilke egenskaber/evner (sentience, bevidsthed, rationalitet osv.) der ”kvalificerer” en entitet, og herunder et AI-system, til moralsk status.
Nærværende afsnit vil præsentere en anden måde at tilgå dette spørgsmål – en tilgang der retter fokus mod de relationer vi indgår i med kunstigt intelligente robotter, og hvilken moralsk betydning disse relationer har. Denne relationelle og sociale tilgang er især forbundet med Mark Coeckelbergh (Coeckelbergh, 2014) og David J. Gunkel (Gunkel, 2012). Nærværende afsnit vil imidlertid udelukkende fokusere på Coeckelbergh.

Mark Coeckelberghs relationelle tilgang

I artiklen The Moral Standing of Machines: Towards a Relational and Non-Cartesian Moral Hermeneutics (2014) argumenterer den belgiske filosof Mark Coeckelbergh for en alternativ tilgang

til spørgsmålet om kunstig intelligens' moralske status. Mere specifikt argumenterer han for, hvad han kalder, ”A relational approach” (Coeckelbergh, 2014, s. 64). Kort sagt, er moralsk status, ifølge denne relationelle tilgang, noget som opstår i relationerne mellem entiteter (Coeckelbergh 2014, s. 64). Hvad dette helt præcist indebærer, vil blive uddybet senere i nærværende afsnit. Først er det formålstjenligt at beskrive, hvad Coeckelbergh kalder ”The standard approach” (Coeckelbergh 2014, s. 62). Der er nemlig, ifølge Coeckelbergh, gode grunde til at udskifte denne med den relationelle tilgang.

”The standard approach”

Moralfilosofiske diskussioner om moralsk status tager, ifølge Coeckelbergh, typisk udgangspunkt i, hvad Coeckelbergh kalder ”the standard approach”. Med dette udtryk refererer Coeckelbergh til en bestemt forståelsesramme, hvori moralsk status forstås som noget, en entitet kan have, hvis den besidder de moralsk relevante egenskaber (Coeckelbergh, 2014, s. 62). Coeckelbergh opstiller denne forståelsesramme på følgende måde:

”Entity x has property p

any entity that has property p, has moral status s

entity x has moral status s” (Coeckelbergh, 2014, s. 63).

Denne måde at forstå moralsk status på er typisk inden for både utilitarismen og deontologien. Hvilke egenskaber der er moralsk relevante er de givetvis uenige i. Pointen er blot, at begge teorier tager udgangspunkt i samme egenskabsbaserede forståelse af moralsk status (Coeckelbergh, 2014, s. 62).

”The standard approach” og dens problematikker

Denne egenskabsbaserede forståelse af moralsk status er ganske vist den mest udbredte, men den er, ifølge Coeckelbergh, uhensigtsmæssig og problemfyldt. Der er to slags problemer: et epistemologisk problem og et problem som Coeckelbergh beskriver som ”The the problem of the gap between reasoning and experience, between thinking and action, between belief and feeling.”

(Coeckelbergh, 2014, s. 63). Hvad denne sidstnævnte problematik helt præcist består i, uddybes efter en kort belysning af det epistemologiske problem.

Det epistemologiske problem består i virkeligheden af to problemer. Det første problem består i, at vi, ifølge Coeckelbergh, er begrænsede i vores evne til rent faktisk at vide, om en given entitet besidder en given egenskab:

”Properties such as “being subject of a life”, “capacity to suffer”, or “consciousness” are notoriously difficult to determine and claims are often contested. Skepticism tells us that we can never be certain about the internal states of other entities. We cannot directly observe properties such as “being subject of a life”. Therefore, it seems difficult to establish the first premise.” (Coeckelbergh, 2014, s. 63).

Denne epistemologiske begrænsning gør det svært at vedligeholde den første præmis, ”entity x has property p”, som ”the standard approach” tager udgangspunkt i (Coeckelbergh, 2014, s. 63).

Andet præmis, ”any entity that has property p, has moral status s”, er, ifølge Coeckelbergh, ligeledes af tvivlsom karakter: Hvordan kan vi vide, at en given egenskab (fx sentience), berettiger/begrunder en given moralsk status? Yderligere er denne tvivl større, når det vedrører kunstigt intelligente robotter, da der er tale om nye entiteter (Coeckelbergh, 2014, s. 63).

Hvis begge præmisser kan betvivles på denne måde, er også konklusionen, ”entity x has moral status s”, tvivlsom. Disse problemer udgør altså, hvad Coeckelbergh anser som et grundlæggende epistemologisk problem (Coeckelbergh, 2014, s. 63).

Der er, som tidligere nævnt, endnu et problem med ”the standard approach”. Denne problemstilling vedrører den uoverensstemmelse/diskrepans, der, ifølge Coeckelbergh, ofte er mellem, på den ene side, en given teoris forståelse af en given entitets moralske status, og på den anden side, vores egentlige erfaring/oplevelse af den pågældende entitet (Coeckelbergh, 2014, s. 63).

Selv hvis vi formåede at nå frem til, i Coeckelberghs ord, en ”moral metaphysics” eller en ”moral science”, hvorudfra en entitets moralske status eller mangel på moralsk status kunne bestemmes, ville vi ofte erfare/opleve entiteten som uforenelig med denne bestemmelse (Coeckelbergh, 2014,

s. 63). Coeckelbergh nævner kunstigt intelligente og autonome robotter som et eksempel på en entitet, vi muligvis kunne erfare/opleve, som noget andet end det, vores teoretiske ramme bestemmer den som. Hvis vi i vores møde med disse robotter erfarer/oplever dem som "more than machines", og måske endda føler en reel forbindelse til dem på trods af, at vores teoretiske ramme ikke tildeler disse robotter nogen moralsk status, er der tale om en uoverensstemmelse mellem vores teoretiske forståelse af disse robotter og vores egentlige erfaring/oplevelse af dem (Coeckelbergh, 2014, s. 63-64).

Denne uoverensstemmelse medfører, at vi, i de tilfælde/situationer, hvor vi måske føler en reel forbindelse til en robot, føler os nødsaget til at beskrive denne forbindelse som irrationel eller barnlig. For hvordan kan vi tage sådanne forbindelser (moralsk) seriøst, hvis vores "moral science" fortæller os, at robotter ikke har nogen moralsk status og ikke er "more than machines"? (Coeckelbergh, 2014, s. 63-64). Det er bl.a. dette spørgsmål, som Coeckelbergh ønsker at adressere med hans alternative relationelle tilgang.

En alternativ tilgang til spørgsmålet om moralsk status

Efter ovenstående afsnit står det klart, at "the standard approach", ifølge Coeckelbergh, er problemfyldt. Der er brug for en anden måde at gribe spørgsmål om moralsk status an. Det er her, at Coeckelberghs relationelle tilgang byder sig til. Som navnet antyder, står relationer som det centrale i denne teori. Coeckelbergh beskriver hans relationelle tilgang på følgende måde:

"I have argued for a relational approach to moral status, which sees moral status as something that emerges through relations between entities . . . Applied to robots, this approach would mean that in order to determine their moral standing, one would need to know its relations with other machines and with humans." (Coeckelbergh, 2014, s. 64).

Moralsk status er altså ikke noget, som en individuel entitet besidder (eller ikke besidder), men noget som opstår i en konkret relation mellem to entiteter. Yderligere skal relationer ikke blot forstås, som relationer mellem to entiteter, men også som entiteternes relation til og ståsted i historiske, sociale, lingvistiske, materielle og kulturelle relationer. Disse relationer påvirker og former den konkrete relation mellem to entiteter (Coeckelbergh, 2014, s. 64-66 & s. 70).

Coeckelbergh pointerer imidlertid, at vi skal passe på med ikke at gøre dette fokus på relationer til endnu en "moral metaphysics". Med dette mener han, at vi skal passe på med ikke at gøre relationer til endnu en metafysisk abstraktion. I så fald ville vi blot udskifte egenskaber (sentience osv.) med relationer, og på den måde gentage eller vedligeholde "the standard approach" (Coeckelbergh, 2014, s. 65).

Vi skal, ifølge Coeckelbergh, ikke diskutere metafysik som det primære fokuspunkt, når det drejer sig om moralsk status. Vi skal derimod rette vores fokus mod den væsentlighed, som epistemologi har i spørgsmål om moralsk status. Vi skal med andre ord stille spørgsmålet: "Which knowledge and what kind of knowledge can we have of entities and of their moral standing?" (Coeckelbergh, 2014, s. 65). For at gennemføre denne epistemologiske drejning skal vi inddrage det moralske subjekt (dvs. os mennesker) i diskussioner om et (moralsk)objekts moralske status. Objekt skal i denne sammenhæng forstås, som den entitet, vi (subjektet) indgår i en relation med (Coeckelbergh, 2014, s. 65). Coeckelbergh kritiserer "the standard approach" for ikke at inddrage subjektet i diskussioner om moralsk status. Ved ikke at tage stilling til subjektet antages det, at moralsk status er noget vi objektivt kan begrunde ud fra en metafysisk egenskab (fx bevidsthed, sentience, rationalitet osv.). På den måde isoleres og abstraheres diskussionen om moralsk status fra den konkrete relation mellem subjekt og objekt, og den konkrete situation, de befinder sig i (Coeckelbergh, 2014, s. 64-65).

Coeckelberghs pointe er, at vi netop skal inddrage subjektet i diskussioner om et objekts moralske status. I den konkrete relation mellem subjekt og objekt skal subjektet tage stilling, til hvordan objektet fremstår for ham/hende (Coeckelbergh, 2014, s. 65-66). Det er, grundet de epistemologiske begrænsninger beskrevet tidligere, ikke muligt at tage stilling til, hvordan objektet "i virkeligheden er" i en metafysisk og objektiv forstand (Coeckelbergh, 2014, s. 63 & 65.) Coeckelbergh beskriver subjektets afgørende rolle i diskussioner om moralsk status på følgende måde:

"Discussions about moral standing take place within language and thinking, more specifically within human language. Humans are the subject of moral status ascription. This implies that the status of the object of moral standing is not independent from human subjectivity. The entity appears to us

in a particular way, and its appearance depends on human subjectivity, that is, on human thinking and also on specific thinking and specific cultures and forms of life at a particular point in history." (Cockelbergh, 2014, s. 65).

Det står altså klart, at spørgsmålet om moralsk status, ifølge Coeckelbergh, ikke kan opstilles på den måde, som "the standard approach" opstiller det på – som et spørgsmål om en given entitets besiddelse af eller mangel på en moralsk relevant egenskab (Coeckelbergh, 2014, s. 65). Det er derimod et spørgsmål om at forholde sig (som subjekt) til den konkrete relation, som man står i med en konkret entitet, som i dens fremtræden henvender sig til subjektet. Det er ud fra denne konkrete relation, at entitetens moralske status opstår (Coeckelbergh, 2014, s. 66). I den forstand er der, hvad Coeckelbergh kalder multisubjektivitet, og en pluralitet af sandheder. Med dette mener han, at en entitets moralske status ikke på forhånd er fastsat, men kan variere alt efter, hvilket moralsk subjekt den indgår i en relation med. På den måde er der ikke en sandhed vedrørende entitetens moralske status, men flere mulige sandheder alt efter hvilken konkret relation den indgår i, og med hvilket konkret subjekt den indgår i relationen med (Coeckelbergh, 2014, s. 66).

Hvilken betydning har denne alternative relationelle tilgang så for kunstigt intelligente robotters moralske status? Det har den betydning, at vi, i stedet for at prøve at bestemme deres moralske status ud fra, hvorvidt de besidder visse moralske relevante egenskaber, skal tage udgangspunkt i konkrete relationer mellem mennesker og kunstigt intelligente robotter. Det er i disse relationer, at et moralsk subjekt kan se robotten overfor sig som et moralsk objekt, frem for bare et objekt (Coeckelbergh, 2014, s. 66). Dette er ikke ensbetydende med, at kunstigt intelligente robotter altid har moralsk status, bare fordi de indgår i en relation med et subjekt. Pointen er blot, at muligheden for denne tildeling af moralsk status altid finder sted i en konkret relation (Coeckelbergh, 2014, s. 67). Coeckelbergh giver følgende eksempel på en menneske-robot relation, hvor robottens moralske status er impliceret:

"Will people start saying, as they tend to say of people who have "met their dog", that someone has "met her robot"? Would such a person, having that kind of relation with that robot, still feel shame at all in front of the robot? And is there, at that point of personal engagement, still a need to talk

about the “moral standing” of the robot? Is not moral quality already implied in the very relation that has emerged here?” (Coeckelbergh, 2014, s. 69-70).

Den konkrete relation som her er beskrevet, sammenlignes med den relation en hundeejer har med sin hund. Der er formentlig tale om en relation, som er gensidig. Dvs., en relation der gavner og bidrager positivt til både personens og robottens tilværelse. Robotten betyder noget for den pågældende person. Den er mere end blot en maskine. I den forstand har robotten moralsk status. Dens moralske status udspringer/opstår af den konkrete relation den har med en konkret person/subjekt, som erfarer/oplever robotten som moralsk betydningsfuld/som et moralsk objekt (Coeckelbergh, 2014, s. 71).

5.5 Adfærd/opførsel som kriterie for moralsk status

Som vi så i det foregående afsnit, kan spørgsmål om moralsk status tilgås på en alternativ måde. Nærværende afsnit vil præsentere en position, som ligeledes må siges at være alternativ. En position, som dens fortaler, John Danaher, kalder ”Ethical Behaviourism”. Som navnet antyder, står adfærd/opførsel centralt i denne tilgang til spørgsmål om moralsk status. Et væsentligt fællestræk mellem denne behavioristiske tilgang og den relationelle tilgang er, at de begge udfordrer egenskabsbaserede tilgange til spørgsmål om moralsk status. Ifølge denne behavioristiske tilgang er det afgørende spørgsmål ikke: Har denne entitet egenskab x? Men: Opfører denne entitet x sig på samme måde som en entitet y, der allerede opfattes som havende moralsk status? (Danaher, 2020, s. 2029-2030).

Den følgende redegørelse omhandler John Danahers behavioristiske position og mere specifikt positionens relevans med hensyn til robotters moralske status. Inden denne dybdegående redegørelse er det hensigtsmæssigt at påpege en mindre, men væsentlig detalje, der muligvis kan forårsage uklarhed eller forvirring hos læseren. I artiklen, som følgende redegørelse tager udgangspunkt i, beskriver Danaher sin position i forhold til robotter, men nævner ikke, om der er tale om kunstigt intelligente robotter. I andre artikler fremstår det imidlertid som om, at Danaher henviser til kunstigt intelligente robotter, når han bruger ordet ”robot” (Danaher, 2019; Sætra & Danaher, 2022). De evner og egenskaber (bevidsthed, sentience, intelligens osv.), som Danaher nævner i den artikel, som redegørelsen tager udgangspunkt i, tyder også på at han med ordet

”robot”, mener kunstigt intelligente robotter. Med denne vigtige detalje in mente kan redegørelsen begynde.

John Danahers ”ethical behaviourism”

I artiklen, Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism (2020), forsvarer John Danaher sin behavioristiske tilgang til spørgsmål om moralsk status, og argumenterer for, at det er ud fra denne tilgang, at robotters moralske status skal bestemmes.

Danaher mener, ligesom Coeckelbergh, at den egenskabsbaserede tilgang til spørgsmål om moralsk status er plaget af epistemologiske problemer. Danaher trækker på Kants opdeling mellem tingen i sig selv og tingen for os til at belyse disse problemer. Vi er epistemisk begrænsede i den forstand, at vi ikke er i stand til at erfare disse moralsk relevante egenskaber (fx sentience eller bevidsthed) i sig selv. Vi kan kun erfare repræsentationer af disse egenskaber. Etisk behaviorisme respekterer vores epistemiske begrænsning og inkluderer den som et vilkår, hvorudfra anvendt etik bedrives. Den behavioristiske tilgang beskæftiger sig derfor med egenskabernes repræsentationer og antager ikke at kunne beskæftige sig med egenskaberne i sig selv. (Danaher, 2020, s. 2028). En vigtig pointe er imidlertid, at en fortaler for etisk behaviorisme ikke udelukker, at disse egenskaber udgør ”the ultimate metaphysical ground for our ethical principles” (Danaher, 2020, s. 2027-2028). Men på grund af vores epistemiske begrænsning må vi forholde os til egenskabernes repræsentation gennem en entitet, og derudfra bestemme entitetens moralske status (Danaher, 2020, s. 2027-2028).

Men hvad er så disse repræsentationer? Ifølge Danaher repræsenteres egenskaberne for os gennem en entitets adfærd/opførsel:

”Many principles concerning the moral status of others depend on metaphysical properties that cannot be directly assessed . . . The ethical behaviourist points out that our ability to ascertain the existence of each and every one of these metaphysical properties is ultimately dependent on some inference from a set of behavioural representations. Behaviour is then, for practical purposes, the only insight we have into the metaphysical grounding for moral status.” (Danaher, 2020, s. 2028).

Det er altså ud fra en entitets adfærd/opførsel, at dens moralske status bestemmes. Dette sker gennem en sammenligning mellem to entiteter. Denne sammenligning forudsætter imidlertid, at der i forvejen eksisterer entiteter, som der er bred enighed om har moralsk status (Danaher, 2020, s. 2040). Sammenligning består så i, at en given entitets adfærd/opførsel sammenlignes med den adfærd/opførsel, som en entitet med moralsk status udviser. Hvis entitetens adfærd/opførsel er magen til den adfærd/opførsel, som entiteten med moralsk status udviser, har førstnævnte entitet den samme moralske status (Danaher, 2020, s. 2029-2030).

To yderligere pointer er vigtige for at forstå denne sammenligningsproces. For det første: hvis en entitet tildeles moralsk status på baggrund af denne sammenligningsproces, er det, fordi den udviser adfærd/opførsel, der ligner den moralsk relevante adfærd/opførsel, som en entitet, vi allerede er enige om har moralsk status, udviser. Der er altså ikke tale om hvilken som helst form for adfærd/opførsel. Der skal være tale om moralsk relevant adfærd/opførsel (Danaher, 2020, s. 2026, s. 2028 & s. 2030). Men hvordan ved vi, hvad der er moralsk relevant adfærd/opførsel? Det er her den anden pointe kommer i spil. Som tidligere nævnt benægter den etiske behaviorisme ikke, at egenskaber som sentience og bevidsthed udgør den metafysiske basis for moralsk status. Derfor tager vi stadig udgangspunkt i disse egenskaber, når vi beskæftiger os med spørgsmål om moralsk status (Danaher, 2020, s. 2029). Den moralsk relevante adfærd/opførsel skal derfor forstås som repræsentationer af den egenskab, som vi mener udgør den metafysiske basis for moralsk status (fx sentience) (Danaher, 2020, s. 2026).

Det oplagte spørgsmål er så, hvor meget en entitets adfærd/opførsel skal ligne denne moralsk relevante adfærd/opførsel. Ifølge Danaher skal der være tale om en ”rough performative equivalence”. Der vil altid være forskelle mellem to entiteter med moralsk status. Tager man udgangspunkt i to mennesker, vil der fx være forskelle i deres respektive karrierevalg, meninger/opfattelser, vaner og udseende. Disse forskelle ændrer ikke på, at de to mennesker har samme moralske status. Det afgørende er, at de, hvad angår den moralsk relevante adfærd/opførsel, er omtrent lignende/ens performativt set. Der er altså ikke et krav om, at de to entiteter skal være identiske (Danaher, 2020, s. 2026).

Denne sammenligningsproces står altså centralt i den etiske behaviorisme. Danaher præsenterer derfor, hvad han kalder ”den etiske behaviorismes komparative princip” (Danaher, 2020, s. 2030). Princippet lyder som følgende:

”If an entity X displays or exhibits roughly equivalent behavioural patterns (P1…Pn) to entity Y, and if it is believed that those patterns ground or justify our ascription of rights and duties to entity Y, then either (a) the same rights and duties must be ascribed to X or (b) the use of P1…Pn to ground our ethical duties to Y must be reevaluated.” (Danaher, 2020, s. 2030).

Det er med udgangspunkt i dette princip, at vi kan bestemme en given entitets moralske status. Dette gælder naturligvis også for robotter (Danaher, 2020, s. 2025). Danahar argumenterer derfor for robotters moralske status på baggrund af ovenstående princip. Danaher bygger argumentet op på følgende måde:

”(1) If a robot is roughly performatively equivalent to another entity whom, it is widely agreed, has signifcant moral status, then it is right and proper to aford the robot that same status. (2) Robots can be roughly performatively equivalent to other entities whom, it is widely agreed, have signifcant moral status. (3) Therefore, it can be right and proper to aford robots signifcant moral status.” (Danaher, 2020, s. 2025).

Danaher pointerer, at der måske allerede er nogle robotter, som kvalificerer sig til moralsk status. Dette afhænger af hvor vi sætter grænsen for, hvad der kvalificeres som ”rough performative equivalence”. Sættes grænsen lavt, er der muligvis allerede robotter, som er berettiget til moralsk status. Danaher påpeger, at mennesker med alvorlige og permanente kognitive og/eller fysiske handicap anses for at have moralsk status. Ligeledes anses dyr med et minimalt adfærdsrepertoire for at have moralsk status. Ifølge Danaher er det ikke utænkeligt, at robotter i den nære fremtid kan tilnærme sig adfærd/opførsel, der ligner disse menneskers eller dyrs adfærd/opførsel. Visse robotter har måske allerede nået dette punkt. I tilfælde af, at dette punkt nås, skal robotterne tildeles samme moralske status som disse mennesker eller dyr (Danaher, 2020, s. 2041-2042).

Danaher nævner endnu et eksempel, som belyser hvordan behaviorismens komparative princip anvendes i spørgsmål om robotters moralske status. Hvis sentience eller evnen til at føle/opleve

smerte er en basis for moralsk status, skal en robot, der regelmæssigt opfører sig eller udviser en adfærd, der indikerer, at den føler smerte, have samme moralske status som enhver anden entitet, hvis moralske status baseres på dens evne til at føle/opleve smerte (Danaher, 2020, s. 2026 & s. 2030).

Danahers etiske behaviorisme står altså i stærk kontrast til de egenskabsbaserede tilgange, der blev beskrevet i tidligere afsnit (fx DeGrazias fokus på egenskaben sentience). I Danahers ord har den etiske behaviorisme det grundlæggende synspunkt, at "what's going on 'on the inside' does not matter from an ethical perspective." (Danaher, 2020, s. 2025). Vi er epistemisk begrænsede. Vi har ikke direkte adgang til disse indre og metafysiske egenskaber. Vi skal derfor forholde os til disse egenskabers repræsentationer. Det gør vi ved at forholde os til entiteters (robotters) adfærd/opførsel (Danaher, 2020, s. 2028).

6. Diskussion – en kritisk gennemgang af positionernes kerneargumenter

Nærværende hovedafsnit er specialets diskuterende del. Her vil de forskellige positioners kerneargumenter diskuteres, modstilles og vurderes. Der vil altså være en kritisk gennemgang af de forskellige positioner. Inden dette er der en kort, men vigtig, pointe, der vedrører forskellen mellem AI-systemer, der er rent software-baserede, og AI-systemer, der er indlejret i hardwareenheder. Størstedelen af den filosofiske litteratur, som dette projekt tager udgangspunkt i, beskæftiger sig specifikt med kunstigt intelligente robotter; dvs. AI-systemer der er indlejret i hardwareenheder. Formålet med nærværende projekt er derimod at beskæftige sig med AI-systemer generelt. Spørgsmålet om kunstigt intelligente systemers moralske status retter sig altså mod både software-baserede AI-systemer, og AI-systemer som er indlejret i hardwareenheder. Som Martin Gibert og Dominic Martin korrekt pointerer, vil rent software-baserede AI-systemer måske ikke fremkalde samme kognitive og emotionelle respons hos mennesker, som et AI-system, der er indlejret i en hardwareenhed (fx en robot), er i stand til (Gibert & Martin, 2022, s. 320). Dette er dog ikke i sig selv et argument for, at det udelukkende er AI-systemer, der er indlejret i hardwareenheder, som kan have moralsk status. Hvis det at have moralsk status er et spørgsmål om at have visse egenskaber (sentience), er det vel irrelevant om entiteten, som har disse egenskaber, er software-baseret eller

om den består af fysisk materiale. Dette er også i overensstemmelse med Boströms og Yudowskys "Principle of Substrate Non-Discrimination" (Boström & Yudowsky, 2011, s. 8).

6.1 Sentience versus bevidsthed – mål og interesser

Som vi så i specialets redegørende del, er sentience og bevidsthed centrale begreber i diskussioner om moralsk status. Dette gør sig også gældende i diskussionen om AI-systemers moralske status. Der er imidlertid langtfra enighed om svarene på nogle af de grundlæggende spørgsmål, der melder sig inden for denne diskussion. Er sentience en nødvendig betingelse for at have moralsk status? Eller er det blot en tilstrækkelig betingelse? Og er bevidsthed overhovedet en nødvendig betingelse for at have moralsk status? Eller er det også blot en tilstrækkelig betingelse?
Spørger man DeGrazia, er sentience en nødvendig og tilstrækkelig betingelse. Bevidsthed er dog en nødvendig betingelse for overhovedet at have sentience (DeGrazia, 2022, s. 74). Neely, Ladak og Shepherd mener derimod, at sentience er tilstrækkeligt, men ikke nødvendigt for at have moralsk status (Neely, 2014, s. 98; Ladak, 2023; Shepherd, 2023, s. 153). Bevidsthed foruden sentience kan også være tilstrækkeligt. Yderligere, kan en entitet som hverken er bevidst eller sentient også have moralsk status, hvis den har mål og interesser (Neely, 2014, s. 103; Ladak, 2023; Shepherd, 2023, s. 152-153).
Netop mål og interesser står som fællesnævneren for disse forskellige positioner: Det at have moralsk status er et spørgsmål om at have interesser, der har moralsk betydning for den pågældende entitets egen skyld. Men giver det overhovedet mening at tale om ikke-sentiente interesser eller mål som havende moralsk betydning? Ifølge Neely, Ladak og Shepherd gør det. De kommer hver især med eksempler, der fungerer som argumenter for, at ikke-sentiente interesser eller mål kan have moralsk betydning.

Som vi så i redegørelsen, eksemplificerer og argumenterer Neely for ikke-sentiente interessers moralske betydning ved at beskrive et AI-system uden hverken sentience eller bevidsthed. Hvis dette AI-system er i stand til at målsætte, uden at disse mål blot består i, at systemet følger en algoritme eller et program, har systemet en grundlæggende form for autonomi. Denne autonomi udgør en tilstrækkelig betingelse for, at systemet har interesser (Neely, 2014, s. 102-103).

Ladak giver et lignende eksempel, der beskriver et skakspillende computerprogram, hvis mål er at vinde. Selvom positive eller negative følelser/oplevelser udebliver, har systemet en præference/et mål: at vinde. Systemet har altså en interesse på trods af dets mangel på sentience (Ladak, 2023).

Shepherd nævner en række eksempler på objektive goder: "the satisfaction of desires, the acquisition of knowledge, the nurturing of relationships of care, the pursuit of long-term plans, and the realization of significant achievements" (Shepherd, 2023, s. 152). En ikke-bevidst entitet som stræber efter disse goder, har moralsk betydelige mål og projekter, som giver den en vis grad af moralsk status (Shepherd, 2023, s. 152-153).

Det er muligt, at entiteterne i disse eksempler kan beskrives som havende mål. Dette forekommer især passende, hvis der er tale om bevidste, men ikke sentiente entiteter. Det er ikke lige så tydeligt, om denne beskrivelse er velvalgt, når der fx er tale om Neelys autonome, men ikke-bevidste og ikke-sentiente computerprogram. Uanset hvad, melder der sig nogle opklarende spørgsmål: Er det mål, som, alt efter om det lykkedes at indfri målene eller ej, er forbundet med positive eller negative følelser/oplevelser? Mål, som kan siges at være i entiteternes interesse at opfylde? Mål, som betyder noget for entiteterne? Dette er sandsynligvis spørgsmål, som DeGrazia ville stille. Hans svar på alle tre spørgsmål ville være nej. Uden sentience er det umuligt for entiteterne at have nogle reelle interesser. Moralske agenter kan ikke handle ud fra disse entiteters egen skyld, da der ikke er nogen interesser der betyder noget for entiteterne (DeGrazia, 2022, s. 76). Uden sentience er der ingen subjektive oplevelser/følelser, som kan påvirkes i positiv eller negativ retning:

"Sentience is the capacity to have pleasant or unpleasant experiences. Any being who can have such experiences has an experiential welfare: subjective experience that can go well or badly from the being's point of view . . . A being or entity that is entirely incapable of having pleasant or unpleasant experiences cannot care what happens to it." (DeGrazia, 2022, s. 77).

Det er svært at se, hvordan mål eller interesser, der er fuldstændigt frakoblet nogen form for positive eller negative følelser/oplevelser, kan betyde noget for den pågældende entitet. Intuitivt forekommer det misvisende overhovedet at tale om disse mål og interesser som entitetens mål og interesser. Hvis indfrielsen af en entitets mål eller interesser ikke er forbundet med positive følelser/oplevelser, og det modsatte ikke er forbundet med negative følelser/oplevelser, giver det

vel ingen mening at tale om disse mål og interesser som egentlige mål og interesser. Har Ladaks skak-spillende computerprogram en reel interesse i at vinde, hvis det at vinde ikke er forbundet med positive følelser/oplevelser, og det at tabe ikke er forbundet med negative følelser/oplevelser? Det er svært at se, hvordan det betyder noget for computerprogrammet, at det vinder frem for at tabe. Det samme kan siges om Neelys og Shepherds eksempler. Alene det at et AI-system er autonomt, forekommer ikke at være tilstrækkeligt, for at det kan have moralsk status. Og yderligere: at et AI-system er autonomt og bevidst, forekommer ikke at være tilstrækkeligt, for at det har moralsk status. Uden sentience fremstår det meningsløst at tale om disse AI-systemer, som entiteter der har interesser. Hvordan kan mål og interesser have nogen betydning for entiteten, hvis ikke det er fordi, at de er forbundet med positive følelser/oplevelser eller negative følelser/oplevelser? Som DeGrazia pointerer, er det at have interesser forbundet med en vis sårbarhed:

”Complete emotional indifference and lack of feeling more generally make them [bevidste entiteter uden sentience] entirely invulnerable such that they lack interests . . . Much of the point of ascribing moral status is to note the vulnerability of certain beings as well as the moral importance of being responsive to their vulnerability. If one is totally invulnerable, on the present view, then one lacks moral status.” (DeGrazia, 2022, s. 82).

AI-systemer, der er bevidste, men uden at have sentience, eller AI-systemer, der hverken er bevidste eller har sentience, har ikke moralsk status. De har ingen reelle interesser, som betyder noget for dem, da de ikke er i stand til at opleve/føle noget som positivt/fornøjeligt eller negativt/lidelsesfyldt. Sentience er en nødvendig og tilstrækkelig betingelse for, at et kunstigt intelligent system kan have moralsk status.

6.2 Den kantianske tilgang

Som vi så i specialets redegørende del, er det muligt at inddrage Kants idé om rationelle personers ubetingede værdighed i diskussionen om AI-systemers moralske status. Ifølge John-Stewart Gordon og David J. Gunkel er dette muligt, fordi Kants begreb om ”personhood” omfatter alle rationelle personer, og personer skal i denne sammenhæng ikke udelukkende forstås som mennesker (Gordon & Gunkel s. 96). Det er ikke usandsynligt, at kunstigt intelligente robotter i fremtiden vil være i stand til at træffe autonome og rationelle beslutninger og dermed handle i overensstemmelse med

moralske principper. I så fald ville de være rationelle personer med ubetinget værdighed (Gordon & Gunkel, 2022, s. 98).

Det åbenlyse spørgsmål er dog, om dette ikke er for høje krav at stille? Ikke blot i forhold til kunstigt intelligente robotter eller andre AI-systemer, men generelt. Hvis avancerede kognitive evner som autonomi og rationalitet er nødvendige betingelser for moralsk status, vil mange af de entiteter, som vi i dag anser for at have moralsk status, ikke opfylde disse betingelser og vil derfor ikke have moralsk status. Netop denne pointe fremhæves af Martin Gibert og Dominic Martin:

There is a good reason to reject intelligence as a criterion for moral status. That is, the property of being intelligent seems totally disconnected from the possibility to be treated morally. Indeed, reaching a threshold of intelligence is generally not required for a human being to have moral status. We usually consider that babies or mentally disabled people can be wronged, even if they do not have the cognitive capacities of a typical human adult (Gibert & Martin, 2022, s. 324).

Gibert og Martin bruger ganske vist ordet ”intelligens”, men de gør det klart, at de bl.a. har begreber som autonomi og rationalitet i tankerne, når de taler om intelligens (Gibert & Martin, 2022, s. 323-324). Gibert og Martin nævner specifikt spædbørn og mentalt handicappede mennesker, som eksempler på entiteter, vi i dag anser som havende moralsk status, selvom de mangler avancerede kognitive evner. Men man kunne også nævne dyr. Fælles for disse eksempler er, at de har sentience. Uanset graden af autonomi og rationalitet (eller en total mangel af disse egenskaber) har de som sentiente entiteter interesser. Heriblandt en grundlæggende interesse i ikke at blive påført smerte. Entiteter med sentience har evnen til at opleve/føle noget som negativt/lidelsesfyldt eller positivt/fornøjeligt. De har altså et perspektiv/subjektive oplevelser/følelser, som kan påvirkes i positiv eller negativ retning. Det er denne evne, som giver en entitet moralsk status – uanset graden af autonomi og rationalitet eller en total mangel derpå. Det samme gør sig gældende med kunstigt intelligente systemer. Et AI-systems moralske status, eller mangel derpå, burde tage udgangspunkt i spørgsmålet: Har dette AI-system sentience? Og ikke: Er dette AI-system rationelt og autonomt?

En vigtig pointe er imidlertid, at avancerede kognitive evner som rationalitet og autonomi ikke er irrelevant i alle moralske henseender. Evner/egenskaber som rationalitet og autonomi er nødvendige betingelser for at være en moralsk agent (Himma, 2009, s. 22-24). Et kunstigt intelligent

system, som har sentience og er autonomt og rationelt, er ikke bare en moralsk patient (har ikke bare moralsk status), men også en moralsk agent. Hvorvidt dette giver AI-systemet en højere grad af moralsk status, er en diskussion i sig selv. Den væsentlige pointe er blot, at sentience i sig selv er nødvendigt og tilstrækkeligt for at et AI-system har moralsk status.

Påstanden om, at autonomi og rationalitet er nødvendige betingelser for moralsk status, synes at blande begrebet moralsk status (eller moralsk patient) og begrebet moralsk agent sammen (Gibert & Martin, 2022, s. 320-321).
Den kantianske tilgang til moralsk status er ekskluderende/underinkluderende. Hvis vi tager udgangspunkt i denne tilgang, betyder det, at en stor del, hvis ikke størstedelen, af alle entiteter med sentience fratages deres moralske status. Dette inkluderer også et fremtidigt AI-system med sentience. Den kantianske tilgang er derfor uhensigtsmæssig, når det drejer sig om moralsk status generelt. Den er derimod relevant, når det vedrører spørgsmål om, hvem der er moralske agenter.

6.3 Den relationelle tilgang

Den relationelle tilgang til spørgsmål om moralsk status, og mere specifikt i forhold til kunstigt intelligente robotter, repræsenterer, som Mark Coeckelbergh selv gør klart, en alternativ og uortodoks position. Men måske er det netop en alternativ tilgang, der er brug for, når det drejer sig om AI-systemers moralske status? Denne tilgang italesætter den praktiske virkelighed, som moralske spørgsmål ikke bare kan abstrahere sig ud af i form af kontekst-uafhængige beskrivelser af metafysiske egenskaber.
Coeckelbergh har en pointe, når han påpeger, at den konkrete relation i en konkret situation mellem to konkrete entiteter ikke bare kan ses bort fra (Coeckelbergh, 2014, s. 66-67). Der er tilfælde, hvor subjektets følelse/oplevelse af en konkret relation med en konkret entitet er i uoverensstemmelse med en distanceret og teoretisk beskrivelse af entitetens moralske status (Coeckelbergh, 2014, s. 66). De fleste mennesker har vel i et eller andet omfang oplevet, at man knytter sig til noget og oplever/føler en vis følelsesmæssig relation, selvom dette, fra et distanceret og kontekst-uafhængigt synspunkt, forekommer irrationelt. Der er altså tilfælde, hvor der er en diskrepans mellem, hvordan en given entitet fremtræder for et konkret subjekt, og hvordan/hvad entiteten, fra et teoretisk og distanceret synspunkt, i virkeligheden er. Coeckelbergh pointerer også den

væsentlige pointe, at vi er begrænsede i vores evne til at vide om en given entitet rent faktisk besidder en given egenskab (fx sentience eller bevidsthed) (Coeckelbergh, 2014, s. 63).
Den relationelle tilgang synes altså at bidrage med væsentlige pointer, der grundlæggende understreger vigtigheden af den praktiske, konkrete og følte virkelighed vi uundgåeligt befinder os i. Men underminerer disse pointer den egenskabsbaserede tilgang til moralsk status? Er den relationelle tilgangs fokus på det enkelte subjekts følelser/oplevelser problematisk når det drejer sig om moralsk status?

Martin Gibert og Dominic Martin argumenterer for, at den relationelle tilgang er en problematisk tilgang til spørgsmål om moralsk status (Gibert & Martin, 2022, s. 323). Hvis moralsk status er noget, der opstår ud af en konkret relation – og med udgangspunkt i det pågældende subjekts følelser/oplevelser af relationen som værdifuld – er der intet fast og objektivt fundament for vores begreb om moralsk status:

”Different persons can experience valuable relationships with different entities for various reasons, some of them being very contingent . . . If the existence of a valuable relationship serves as an enabling condition for moral status, then moral status becomes very subjective, contingent, or even arbitrary, and cannot be easily universalized . . . The relational argument raises issues of consistency and objectivity.” (Gibert & Martin, 2022, s. 323).

Kestutis Mosakas deler samme bekymring (Mosakas, 2021). Hvis moralsk status ikke er et spørgsmål om at have visse egenskaber, men blot et spørgsmål om det enkelte subjekts følelser over for en entitet, bliver moralsk beslutningstagen blot til relative vurderinger over hvilke entiteter der skal tages moralsk hensyn til:

”Their view [den relationelle tilgang] seems to incur the problem of extreme meta-ethical relativism . . . What the relational approach is fundamentally concerned with is our feelings and attitudes towards different entities, since that is what constitutes the basis of our relations . . . Without any central moral properties or guiding principles, it is difficult to see how this approach could genuinely help us in our moral decision-making without getting bogged down in a sea of relative judgements.” (Mosakas, 2021, s. 434).

Giberts, Martins og Mosakas' argumenter er stærke. De belyser en grundlæggende svaghed i den relationelle tilgang. Hvis en entitets moralske status afhænger af, om den indgår i en relation med et subjekt, som værdsætter den, så er resultatet vel en form for moralsk relativisme.
Som Gibert og Martin pointerer: Hvad hvis en af disse relationer på den ene eller anden måde brydes? Mister entiteten, hvis moralske status afhænger af relationen, så sin moralske status? Dette fremstår kontraintuitivt (Gibert & Martin, 2022, s. 323). En egenskabsbaseret tilgang lider ikke af samme svaghed, da der er et fundament i form af en eller flere egenskaber, der fungerer som en objektiv målestok for, hvilke entiteter der har moralsk status. Coeckelbergh kunne komme med det modargument, at vi ikke er i stand til at vide, om en entitet har en af disse metafysiske egenskaber (Coeckelbergh, 2014, s. 63).Men som Gibert og Martin påpeger, er denne epistemologiske begrænsning ikke i sig selv nogen grund til at droppe den egenskabsbaserede tilgang. Bare fordi vi har svært ved at opnå sikker viden om tilstedeværelsen af disse egenskaber i en entitet, betyder det ikke, at disse egenskaber ikke udgør de rette kriterier/betingelser for moralsk status (Gibert & Martin, 2022, s. 328).

Al denne kritik er dog ikke ensbetydende med, at Coeckelberghs pointer er moralsk irrelevante i alle henseender. Konkrete relationer og de følelser/oplevelser disse indebærer, har moralsk betydning. Hvis fx jeg oplever en følelsesmæssig forbindelse til en AI-robot og på den måde føler, at jeg indgår i en relation med robotten, har robotten muligvis moralsk betydning i den forstand, at jeg har en interesse i, at den fx ikke bliver ødelagt. Dette betyder ikke, at robotten har moralsk status. Dette udelukker dog ikke, at den kan have moralsk betydning, da dens vedligeholdelse er af interesse for en entitet (mig) med moralsk status. En vigtig pointe dertil er, at det selvfølgelig ville være en helt anden situation, hvis AI-robotten rent faktisk havde sentience. I så fald ville robotten have sine egne interesser og dermed moralsk status uafhængigt af, om det er i min interesse at den ikke bliver ødelagt.

En relationel tilgang til spørgsmålet om kunstigt intelligente systemers moralske status er altså uhensigtsmæssig. Uden nogen form for objektiv målestok i form af egenskaber er resultatet en moralsk relativisme, der efterlader moralske spørgsmål i en tilstand af subjektive følelser og holdninger.

6.4 Danahers etiske behaviorisme

John Danahers etiske behaviorisme repræsenterer, ligesom Coeckelberghs relationelle tilgang, også en alternativ tilgang til spørgsmål om moralsk status. Fælles for disse tilgange er deres skeptiske syn på egenskabsbaserede tilgange (Danaher, 2020, s. 2027-2028; Coeckelbergh, 2014, s. 63).
I denne sammenhæng er de ligeledes fælles om et fokus på, hvordan entiteter, herunder AI-robotter, fremtræder for os. Siden vi er epistemisk begrænsede og ikke har direkte adgang til metafysiske egenskaber som fx sentience, er vi nødsaget til at forholde os til, hvordan entiteter, vi indgår i relationer med, fremtræder for os (Coeckelbergh, 2014, s. 65), eller hvordan de moralsk relevante egenskaber fremtræder for os gennem en given entitets adfærd/opførsel (Danaher, 2020, s. 2028). Der synes imidlertid at være den forskel, at Cockelbergh fokuserer på, hvordan entiteter som fx en AI-robot fremtræder for os (Coeckelbergh, 2014, s. 65), hvorimod Danaher fokuserer på hvordan egenskaber som fx sentience fremtræder for os (Danaher, 2020, s. 2028). Når Coeckelbergh beskriver en robots fremtræden for os i en relation, er det umiddelbart uden nogen yderligere bemærkninger om, hvorvidt denne fremtræden er forbundet med nogen af de egenskaber, der typisk tillægges en moralsk relevans (fx bevidsthed, sentience, autonomi, rationalitet osv.) (Coeckelbergh, 2014, s. 65-67 & s. 74). Praktisk ligner disse Coeckelberghs og Danahers tilgange hinanden, da Danaher pointerer, at metafysiske egenskaber fremtræder for os gennem en entitets adfærd/opførsel (Danaher, 2020, s. 65).
I den forstand er spørgsmål om moralsk status for begge teoretikere, et spørgsmål der ikke kan adskilles fra den konkrete og praktiske virkelighed, hvori en entitet fremtræder for os på en bestemt måde. Teoretisk er den beskrevne forskel dog tydelig, da Danaher ikke fuldstændigt opgiver at tage stilling til de egenskaber, der typisk tillægges en moralsk relevans (Danaher, 2020, s. 2027-2028). Danahers etiske behaviorisme bevarer i den forstand en vis grad af objektivitet. Modsat Coeckelbergh, hvor det fremstår som om, at moralsk status afhænger af et subjekts personlige følelser/oplevelser af en given entitet, begrænser Danaher sig til moralsk relevante egenskabers fremtræden.

Men kan det virkelig passe, at vi, på grund af vores epistemiske begrænsning hvad angår metafysiske egenskaber, udelukkende skal forholde os til disse egenskabers fremtræden gennem entiteters opførsel/adfærd, når vi diskuterer eller bestemmer moralsk status? Kan man fx ikke forestille sig en

kunstigt intelligent robot, der er blevet udviklet/designet til at efterligne, at den mærker smerte, selvom den hverken besidder sentience eller bevidsthed? Netop disse spørgsmål stiller Jilles Smids i hans kritik af Danahers etiske behaviorisme (Smids, 2020, s. 2859-2860).

Det er ikke indlysende, hvorfor det udelukkende er en entitets adfærd/opførsel, der fungerer som bevis for dens besiddelse af en moralsk relevant egenskab:

”Once we recognize that we look at the robot’s behaviour with the aim to infer whether or not the robot has the metaphysical property that we believe grounds moral status, why not allow all evidence that might be of relevance to making a justifed inference?” (Smids, 2020, s. 2859).

Når det drejer sig om kunstigt intelligente robotters adfærd/opførsel, forekommer det især at være vigtigt, at man forholder sig skeptisk over for dens adfærd/opførsel. Som Smids pointerer, er det faktum, at vi har at gøre med en robot, en åbenlys årsag til at lede efter andre beviser som kan støtte eller modbevise adfærdens/opførslens moralske relevans (Smids, 2020, s. 2859-2860).
Det Smids hentyder til, er naturligvis at robotter fremstilles og designes (Smids, 2020, s. 2859-2060). Smids giver et eksempel, hvor han sammenligner den adfærd/opførsel, som en robot-hund og en biologisk hund udviser. I eksemplet antages det, at sentience er en egenskab, der tildeler en entitet en betydelig grad af moralsk status. Hvis vi ser en biologisk hund blive slået, og hunden reagerer på dette ved at udvise typisk adfærd/opførsel, der indikerer at den mærker smerte (fx at den hyler og kryber sig sammen), vil den bedste forklaring på denne adfærd/opførsel være, at hunden har sentience. Derfor har hunden en vis grad af moralsk status (Smids, 2020, s. 2859-2860).
Hvis vi derimod observerer den samme situation, men med den ændring, at det er en robot-hund, der udviser adfærden/opførslen, vil den bedste forklaring på denne adfærd/opførsel ikke være, at hunden har sentience og dermed moralsk status (Smids, 2020, s. 2859-2860). Den bedste og mest plausible forklaring ville være, at robot-hunden er designet således, at den udviser denne adfærd/opførsel, når den bliver slået (Smids, 2020, s. 2860). Ifølge Danahers komparative princip har robot-hunden samme moralske status som den biologiske hund. Robot-hunden udviser adfærd/opførsel, der ligner den moralsk relevante adfærd/opførsel, som en entitet (den biologiske hund), vi allerede er enige om har moralsk status, udviser (Danaher, 2020, s. 2030). Men som Smids

eksempel viser, kan dette komparative princip ikke stå alene og ubetinget som bevis for sentience og moralsk status:

”The artifcial nature of the dog separates what is intimately connected in real dogs: having inner pain sensations and exhibiting pain behaviour. It is characteristic . . . For robotic design that it is possible, at least up to a certain extent, to realize behavioural equivalence without having to design a mental life equivalent to that of a real dog . . . For that very reason, behavioural equivalence does not unconditionally support an inference to the presence of mental life. We see that in case of the robot dog, behavioural evidence is not decisive” (Smids, 2020, s. 2860).

Danahers komparative princip er problematisk, da det er overinkluderende i forhold til hvilke entiteter, der har moralsk status. Entiteter der udelukkende udviser en adfærd/opførsel, som indikerer at de har sentience, fordi de er blevet designet til at efterligne denne adfærd/opførsel, skal, ifølge det komparative princip, have samme moralske status som den entitet der efterlignes. Dette resulterer i absurde situationer, hvor entiteter, der tydeligvis ikke har sentience, skal behandles som om de har sentience.

7. Konklusion

Her vil redegørelsen og diskussionen sammenfattes, og problemformuleringen vil blive besvaret:

Hvilke kriterier skal et kunstigt intelligent system opfylde for at have moralsk status?

Det kan konkluderes, at det at have sentience står som det væsentlige kriterie, et kunstigt intelligent system skal opfylde, for at have moralsk status. Evnen til at opleve/føle noget som positivt/fornøjeligt eller negativt/lidelsesfyldt er en nødvendig og tilstrækkelig betingelse for, at et AI-system kan have interesser, og mere specifikt, interesser der betyder noget for AI-systemet. Moralske agenter skal i deres interaktion med/behandling af et AI-system med sentience, tage hensyn til eller medregne disse interesser.

AI-systemer, der er bevidste, men uden at have sentience har ikke nogen reelle interesser, der betyder noget for dem. Det samme gør sig gældende med AI-systemer, der hverken er bevidste eller

har sentience. Disse AI-systemer er ikke i stand til at opleve/føle noget som positivt/fornøjeligt eller negativt/lidelsesfyldt. Derfor har de ingen moralsk status.

Avancerede kognitive evner som autonomi og rationalitet er hverken nødvendige eller tilstrækkelige betingelser for, at et AI-system kan have moralsk status. Det der betyder noget, er om AI-systemet har interesser, som betyder noget for systemet. Dette kræver ikke rationalitet eller autonomi, men sentience. Avancerede kognitive evner er derimod nødvendige betingelser for, at et AI-system kan være en moralsk agent.

Den relationelle tilgang til spørgsmål om moralsk status lider af den grundlæggende svaghed, at den befordrer/fremmer en moralsk relativistisk tilgang til spørgsmål om moralsk status og derfor til moralske spørgsmål generelt. Vi skal ikke bestemme et AI-systems moralske status ud fra den enkeltes subjektive følelser over for AI-systemet. Vi skal derimod tage udgangspunkt i en egenskabsbaseret tilgang (med sentience som den moralsk relevante egenskab).

Danahers etiske behaviorisme er, ligesom den relationelle tilgang, en alternativ tilgang til spørgsmål om moralsk status. Den lider dog ikke af samme svaghed (moralsk relativisme), da den bevarer en vis grad af objektivitet. Danahers komparative princip har imidlertid den konsekvens, at vi ender med at tildele moralsk status til entiteter, der åbenlyst ikke besidder den moralsk relevante egenskab.

Litteraturliste

Anderson, M., & Anderson, S. L. (2007). Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine, 28*(4), 15-26. Retrieved from https://www.proquest.com/scholarly-journals/machine-ethics-creating-ethical-intelligent-agent/docview/208129690/se-2

Bartneck, C., Lütge, C., Wagner, A. R., & Welsh, S. (2021). An introduction to ethics in robotics and AI. In *SpringerBriefs in ethics*.

Boström, N., & Yudowsky, E. (2011), Nickbostrom.com, *The Ethics of Artificial Intelligence*, Retrieved September 6, 2023, from: https://nickbostrom.com/

Coeckelbergh, M. (2014). The Moral Standing of Machines: Towards a Relational and Non-Cartesian Moral Hermeneutics. *Philosophy & Technology*, *27*(1), 61–77. https://doi-org.zorac.aub.aau.dk/10.1007/s13347-013-0133-8

Danaher, J. (2020). Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. *Science and Engineering Ethics*, *26*(4), 2023–2049. https://doi-org.zorac.aub.aau.dk/10.1007/s11948-019-00119-x

Danaher, J. (2019). The rise of the robots and the crisis of moral patiency. *AI & SOCIETY*, *34*(1), 129–136. https://doi-org.zorac.aub.aau.dk/10.1007/s00146-017-0773-9

DeGrazia, D. (2008). Moral Status As a Matter of Degree?. *The Southern Journal of Philosophy*, *46*(2), 181-198. https://doi-org.zorac.aub.aau.dk/10.1111/j.2041-6962.2008.tb00075.x

DeGrazia, D., & Millum, J. (2021). *A Theory of Bioethics*. Cambridge: Cambridge University Press. doi:10.1017/9781009026710

DeGrazia, D. (2022). Robots with Moral Status? *Perspectives in Biology and Medicine, 65*(1), 73-88. https://doi.org/10.1353/pbm.2022.0004

European Commission, *A definition of Artificial Intelligence: main capabilities and scientific disciplines*, (2018), Retrieved September 15, 2023, from: https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

Faktalink, (2023), *Kunstig intelligens*. Retrieved September 15, 2023, from https://faktalink.dk/titelliste/kunstig-intelligens

Gibert, M., & Martin, D. (2022). In search of the moral status of AI: why sentience is a strong argument. *AI & SOCIETY*, *37*(1), 319–330. https://doi-org.zorac.aub.aau.dk/10.1007/s00146-021-01179-z

Gordon, J., & Nyholm, S. (u.å.). Ethics of Artificial Intelligence. *The Internet Encyclopedia of Philosophy*. Retrieved from https://iep.utm.edu/ethics-of-artificial-intelligence/

Gordon, J., & Gunkel, D. J. (2022). Moral status and intelligent robots. *Southern Journal of Philosophy*, *60*(1), 88–117. https://doi-org.zorac.aub.aau.dk/10.1111/sjp.12450

Gruen, L. (2003). The Moral Status of Animals. *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/archives/sum2021/entries/moral-animal/

Gunkel, D. J. (2012). *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. MIT Press.

Gunkel, D. J., & Bryson, J. J. (2014). Introduction to the special issue on Machine Morality: The machine as moral agent and patient. *Philosophy & Technology*, *27*(1), 5–8. https://doi-org.zorac.aub.aau.dk/10.1007/s13347-014-0151-1

Hauser, L. (u.å.). Artificial Intelligence. *The Internet Encyclopedia of Philosophy*. Retrieved from https://iep.utm.edu/artificial-intelligence/#SSH4c.ii

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, *11*(1), 19–29. https://doi-org.zorac.aub.aau.dk/10.1007/s10676-008-9167-5

Jaworska, A. & Tannenbaum, J. (2013). The Grounds of Moral Status. *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/archives/spr2023/entries/grounds-moral-status/

Kamm, F. (2007). *Intricate Ethics: Rights, Responsibilities, and Permissible Harm.* Oxford University Press.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Kant, I. (1998). Groundwork of the Metaphysics of Morals. I M. Gregor, (Red.), *Immanuel Kant: Groundwork of the Metaphysics of Morals* (s. 1-67). Cambridge University Press 1997.

Kubat, M. (2021). *An Introduction to Machine Learning*. (3. udg.). Springer, Cham. https://doi-org.zorac.aub.aau.dk/10.1007/978-3-030-81935-4

Ladak, A. (2023). What would qualify an artificial intelligence for moral standing? *AI And Ethics*. https://doi-org.zorac.aub.aau.dk/10.1007/s43681-023-00260-1

Learned-Miller, E. G. (2014). Introduction to supervised learning. *I: Department of Computer Science, University of Massachusetts*, *3*.

Lindskov, T., Ibrahim, R., Thomsen, M., Bell, B., & Kruse, C. (2020). Kunstig intelligens i det danske sundhedsvæsen. *Ugeskrift for Læger, 182*(V09190540), 1-6.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, Retrieved from: https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

Mosakas, K. (2021). On the moral status of social robots: considering the consciousness criterion. *AI & SOCIETY*, *36*(2), 429–443. https://doi-org.zorac.aub.aau.dk/10.1007/s00146-020-01002-1

Müller, V. C. (2020). Ethics of Artificial Intelligence and Robotics. *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/

Neely, E. L. (2014). Machines and the moral community. *Philosophy & Technology*, *27*(1), 97–111. https://doi-org.zorac.aub.aau.dk/10.1007/s13347-013-0114-y

Singer, P. (1993). *Practical Ethics*. (2. udg.). Cambridge University Press.

Shepherd, J. (2023). Non-Human Moral Status: Problems with Phenomenal Consciousness. *Ajob Neuroscience*, *14*(2), 148–157. https://doi-org.zorac.aub.aau.dk/10.1080/21507740.2022.2148770

Smids, J. (2020). Danaher's Ethical Behaviourism: an adequate guide to assessing the moral status of a robot? *Science and Engineering Ethics*, *26*(5), 2849–2866. https://doi.org/10.1007/s11948-020-00230-4

Sætra, H. S., & Danaher, J. (2022). To each technology its own ethics: the problem of ethical proliferation. *Philosophy & Technology*, *35*(4). https://doi-org.zorac.aub.aau.dk/10.1007/s13347-022-00591-7