Robotters moralske status

Fortrolig Ikke fortrolig X

Prøvens form (sæt kryds)	Projekt	Synopsis	Artikel	Speciale X	Skriftlig opgave

Uddannelsens navn	Kandidat - Anvendt Filosofi	
Semester	10. semester	
Prøvens navn/modul (i studieordningen)	Kandidatspeciale	
Gruppenummer	Studienummer	Underskrift
Navn: Samantha Kjær Jørgensen	**20177346**	
Afleveringsdato	08-08-2022	
Projekttitel/Synopsis titel/Speciale-titel/ opgave nummer	Robotters moralske status	
I henhold til studieordningen må opgaven i alt maks. fylde antal tegn	192.000	
Den afleverede opgave fylder (antal tegn med mellemrum i den afleverede opgave) (indholdsfortegnelse, litteraturliste og bilag medregnes ikke)	120.781	
Vejleder (projekt/synopsis/speciale)	Antje Gimmler	

Jeg/vi bekræfter hermed, at dette er mit/vores originale arbejde, og at jeg/vi alene er ansvarlige for indholdet. Alle anvendte referencer er tydeligt anført. Jeg/vi er informeret om, at plagiering ikke er lovligt og medfører sanktioner. Regler om disciplinære foranstaltninger over for studerende ved Aalborg Universitet (plagiatregler): http://www.plagiat.aau.dk/regler/

Indholdsfortegnelse:

Abstract:

The concept of conscious humanoid robots has been a discussed topic for a long time, which have been current both within philosophy and outside of it. The idea of humanoids achieving consciousness have consistently been mentioned in media and discussed by people, and it has shown to create a great concern for some, while being a victory for others. When comparing how the media portrays humanoid robots compared to how the developers describe them, differences become apparent. The same widely different opinions can be found in the general public, and this dissonance is what I will be diving into in this paper. Humanoid artificial technology has been portrayed very differently throughout games, movies, books, and within topics like morality and ethics. The general viewpoint within the media in based upon the fear of a potential uprising, which results in the fall of humanity. The most seen example of humanoids in books, movies and games is the uprising, since the humanoid robots perceive humans as inferior, but there are also other examples, where the uprising stems from mistreatment and poor living standards. The general viewpoint on robots can already be seen today, where they are perceived as tools, and not even close to humans in moral status. This could be a major ethical and moral issue for us in the future, but also, in the potential event of an artificial intelligence reaching substantial conscience, for the robots. It could be a future issue since a conscious humanoid would not be treated by the moral and ethical standards of conscious beings since the general perception of them is an empty shell. This current behaviour might lead to serious and catastrophic issues in the future, since humans currently do not know how to define consciousness, and therefore it cannot be detected in humanoid robots. This would result in moral and ethical mistreatment of a conscious being. These moral and ethical concerns for the robots will be investigated in this paper and through the analysis of five different philosophers, and I will discuss these issues and suggest potential solutions as well as dive deeper into what the moral status of such a being might entail. This paper will outline potential moral and ethical concerns for the humanoid robots both within the paper and outside of them and develop a potential prevention strategy.

Indledning:

Når der bliver diskuteret i forhold til humanoide robotter så er det som regel fra menneskers synspunkt, hvor fokusset omhandler hvordan etiske problemstillinger kan opstå for mennesker når de eksisterer sammen med humanoide robotter. Selvom disse problemstillinger kan være vigtige i et sådant scenarie, så vil denne opgave ikke omhandle disse problemstillinger. I stedet vil denne opgave prøve at skabe en dialog om en sameksistens mellem mennesker og humanoide robotter og hvilke problemer der kan opstå i forhold til sociale interaktioner mellem mennesker og humanoide robotter. Jeg vil derfor, for at give opgaven validitet formulere mig med henblik på at humanoide robotter kan opnå fænomenal bevidsthed og sprog, og at de vil kunne opnå en bevidsthed og sprog som er på et lignende niveau som mennesker, og derved også kunne opnå samme moralske status. Definitionen i opgaven omhandlende humanoid kunstig intelligens vil være betinget i det humanoide aspekt. Den kunstige intelligens vil i alle aspekter være menneskelig, både i krop og forstand. Der vil være tale om en robot som ligner et menneske både visuelt, intellektuelt, følelsesmæssigt og moralsk. Dette er valgt på baggrund af at meget litteraturen indenfor feltet omkring robotter mangler denne vinkel, men dette er stadig en yderst vigtig vinkel, hvis man på sigt ønsker at undgå umoralske og uetiske handlinger mod robotter. Jeg vil ud fra dette standpunkt forsøge at lokaliserer problemstillinger i forhold til hvordan humanoid kunstig intelligens bør blive behandlet moralsk og etisk set. Jeg vil undersøge dette med en gennemgang af Seibt, Damholdt & Vestergaard, Coeckelbergh, Müller, Schneider og Liao, samt en kort opsummering af kunstig intelligens' historie, Turing testen, og basisviden omkring at arbejde med kunstig intelligens. Disse vil blive undersøgt individuelt i et analyse-/diskussionsafsnit, hvor jeg efterfølgende vil yderligere beskrive og diskutere de problemstillinger som ikke er nævnt i disse tekster. Jeg har begrænset min opgave da jeg fokuserer på humanoide robotter hvor kunstig intelligensaspektet er inkorporeret.

Problemformulering:

Hvilken moralsk status har en humanoid kunstig intelligens og hvordan påvirker denne status de etiske relationer mellem mennesker og den kunstige intelligens?

Redegørelse:

Beskrivelse af teknologien

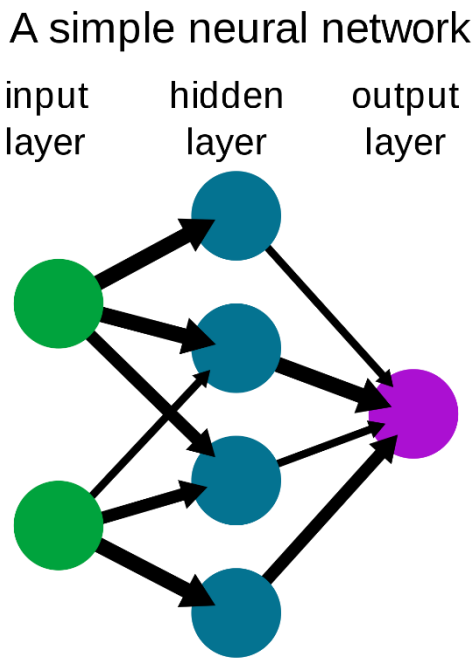
Den officielle start af feltet kunstig intelligens var i 1956 hvor det også blev navngivet. Kunstig intelligens som koncept kan blive fundet længere tilbage, da der er spor tilbage i 1600-tallet, hvor Descartes udtaler sig et koncept, som minder meget om det vi kender som kunstig intelligens i dag. (Bringsjord & Govindarajulu, 2020)

Turing testen blev skabt i 1950 af Alan Turing. Formålet med testen er teste det lingvistiske niveau af en kunstig intelligens, og om hvor vidt der kan skelnes mellem niveauet på et menneskes. Udførslen af testen består af to personer, hvor den ene er deltager og den anden dommer, og personerne er placeret i hvert deres rum. Dommeren stiller de to deltager, hvor den ene er en kunstig intelligens og den anden en person, spørgsmål gennem e-mail. Dommeren skal herefter kunne gætte hvilken e-mail er sendt fra en kunstig intelligens, og hvis svarene ikke er bedre end 50/50 i forhold til hvilke svar er fra den kunstige intelligens, så er Turing testen bestået af den kunstige intelligens. (Bringsjord & Govindarajulu, 2020)

Alan Turing foreslog at tog en anden fremgangsmåde til kunstig intelligens. Man kunne prøve at efterligne den menneskelige fremgang i opbygning af sprogligt niveau, som man ser hos børn. Denne 'børne' kunstige intelligens vil kunne lære sprog langsomt over tid ligesom børn, og naturligt opnå et sprog på niveau med voksne. Turing selv mente at hans test ville blive klaret af en kunstig intelligens inden år 2000, men stadig i dag er det uklart om en kunstig intelligens har bestået testen. (Bringsjord & Govindarajulu, 2020)

Neurale netværker er en del af maskinlæring, som sigter efter at efterligne den menneskelige hjerne. Dette gør den gennem algoritmer, som forsøger at lære på en måde lignende menneskers. Neurale netværker er et system af funktioner, som bearbejder et input, og giver et output baseret på de underliggende funktioner. Et eksempel på et input kan være billeder, hvor formålet er at finde og identificere et objekt i et billede. Artificielle neurale netværker bygger på menneskers viden om den menneskelige hjerne, og hvordan neuroner bruger til at forstå information, som mennesker modtager gennem sanserne. Neurale netværker prøver derfor at genkende mønstre, ligheder og forstå information gennem inputs på en måde som efterligner den menneskelige hjerne og menneskers biologi. (Dhingra, 2021)

Som tidligere beskrevet så prøver et neuralt netværk at efterligne den menneskelige hjerne ved at bruge neuroner. (Dhingra, 2021) I et neuralt netværk så er der som sagt et input, som bliver sendt til et lag i det neurale netværk. Et lag består af neuroner, som har forskellige vægte, som er en indikator for hvor stor betydning den information det enkelte neuron har lært. Den lærte information vil til sidst blive sendt til output laget, som behandler vægtene og derigennem for et resultat. Forskellen på et neuralt netværk og et dybt neuralt netværk er mængden af lag, som er mellem input og output lagene. Hvis der er mere end et lag mellem input og output laget, så ses det som at være deep-learning. (Reyes, 2022)



Figur 1: Et simpelt neuralt netværk (Deepanshi, 2021)

Når man træner neurale netværker, så bruger man to forskellige læringsmetoder, som er supervised og unsupervised læring. Når man snakker om supervised læring, så kræver det et stort forberedelsesarbejde, af den data som man skal bruge til at lære den kunstige intelligens noget. Denne læringsmetode kræver at mindst et individ skal kigge igennem data for at den kunstige intelligens kan opnå gode præcise resultater. Her vil man gennemgå data og markerer og definerer hvad det er som den kunstige intelligens skal lære. Det annoterede data vil man så bruge til at træne sin kunstige intelligens, samt man vil også kunne bruge det til at teste hvor præcis den kunstige intelligens man har lavet er. (Seldon, 2021)

Unsupervised læring derimod er trænet på ikke annoteret data, hvilket betyder at hvor den kunstige intelligens i supervised læring er fortalt hvad den skal lære, så lære den selv i unsupervised læring. Gennem den data ens kunstige intelligens er give, så vil den selv finde sammenhænge og betydninger. Dette er en meget hurtigere metode til at træne ens kunstige intelligens, da det ikke kræver et stort annoteringsarbejde, men det betyder også, at man ikke har kontrol over hvordan ens kunstige intelligens lære. Denne metode er også god til at finde sammenhænge og grupperinger i den data som man bruger, da ens kunstige intelligens naturligt vil arbejde i form af grupperinger. (Seldon, 2021)

Når man træner og udvikler kunstig intelligens, så kan må opnå gode resultater. Men når man snakker om kunstig intelligens, så bliver man også nødt til at nævne black box problemet. Black box problemet opstår primært i forbindelse med neurale netværker hvor man bruger deep-learning, dette er da man kender til inputtet og outputtet af ens kunstige intelligens, men står med intet kendskab til hvordan den kunstige intelligens kom fra A til B. Som tidligere nævnt består deep-learning neurale netværker af mange gemte lag, hvilket kan være flere hundrede eller tusinde, og hvert enkelt lag lære noget forskelligt, med tiden vil det bliver så kompliceret at det er umuligt at hvad de enkelte neuroner har lært. Dette problem bliver også set i sammenhæng med nogle af de mysterier der er omkring den menneskelige hjerne, da man ikke præcist ved hvordan hjernen virker, da der er for mange neuroner til at man kan finde ud af hvad der præcist sker. Siden vi har dette problem med den menneskelige hjerne, så vil man opnå det samme problem med dybe neurale netværker, som prøver at efterligne de processorer som den menneskelige hjerne gennemgår med neuroner. (ThinkAutomation)

Kunstig intelligens er en hurtig voksende industri og bliver brugt af flere og flere virksomheder, da det er blevet mere tilgængeligt, da hardware er blevet tilgængeligt for en overkommelig pris. Der ses stadig en stor forskel i hvor effektive kunstig intelligens er baseret på hvor god hardware det bliver brugt til det. Siden teknologien har gjort kunstig intelligens mere tilgængeligt, så er der sket endnu mere forandring, hvor dybe neurale netværker får flere lag, der bliver brugt større datamængder, og man ønsker stadig en kunstig intelligens, som har en hurtig reaktions tid. Vi begynder at nå et problem igen, hvor at kunstig intelligens bevæger sig for hurtigt, så der ikke er tilgængelig hardware, som kan bruges til det. Dette vil være et at de store problemer i det næste årti, hvor alle ønsker bedre og hurtigere kunstig intelligens, men som kræver det nyeste og dyreste hardware. Der vil derfor på sigt komme en periode, hvor at udviklingen af kunstig intelligens når sit maksimum indtil en ny opdagelse indenfor hardware sker, hvorefter at kunstig intelligens igen ville kunne skabe fremskridt. (Rodriguez, 2021)

Coeckelbergh

"Can we trust robots?" fra 2011 af Mark Coeckelbergh er omkring tillid til teknologi, hvor den specifikt udforsker den øgede brug af robotter i samfundet og deres interaktioner med mennesker. I begyndelsen af teksten specificerer Coeckelbergh at det specifikt er robotter brugt indenfor områderne underholdning, sex, sundhedsfeltet og militært brug. (s. 53-54)

Coeckelbergh (2011) omtaler en eksisterende tillid som allerede eksisterer mellem mennesker og teknologi, hvilket over tid er opstået gennem telefoner, computere og biler. Ud over de redskaber som de fleste bruger i deres hverdag, så bruges der også fjernstyret robotter indenfor forskellige områder som militær, medicin, og de fleste flyvemaskiner har en form for autopilot. Denne tillid vil også de næste år potentielt blive yderligere udvidet, da autonome biler er et nutidigt stort emne, da der sker en stor udvikling, så det ikke længere virker som et uopnåeligt mål. (s. 53)

"I argue that in so far as robots appear human, trusting them requires that they fulfil demanding criteria concerning the appearance of language use, freedom, and social relations" (Coeckelbergh, 2011, s. 54)

De tre nævnte kriterier udgør en stor del af indholdet af denne redegørelse og der vil være en mere dybdegående forklaring senere i redegørelsen. Coeckelbergh (2011) begrunder herigennem at robotterne der diskuteres i teksten, er humanoid, og derved at den omtalte form for tillid der redegøres for, tager udgangspunkt i humanoid robotter. Coeckelbergh (2011) nævner derudover at selvom robotterne ikke har samme form som mennesker, så vil der stadig være grundlag for at diskuterer tilliden, som mennesker har til dem. (s. 54) De tre tidligere nævnte kriterier vil blive beskrevet og forklaret mere i dybden senere i denne del.

Coeckelbergh (2011) bruger dette til at omtale den menneskelige tillid til de nutidige robotter (som han kalder artefakter). Den nuværende tillid beskriver han som "trust as reliance", som tager udgangspunkt i at mennesker har en forventning at robotter udfører en given opgave, dette kunne være at en automatisk maskine støvsuger i huset. Denne omtalte tillid vil dog ikke falde på robotten, da den nærmere vil på falde producenten af støvsugeren i stedet for støvsugeren selv. (s. 54)

Den tillid som eksisterer mellem mennesker beskrives gennem to forskellige former, navnlig kontraktuelindividualistiske form og fænomenologisk-sociale form. Den kontraktuelle-individualistiske vinkel beskrives som individer der indgår i en form for social relation, hvor der derigennem vil der opstå sociale forventninger. Dette er grunden til tilliden mellem de forskellige individer, da mennesker har en vis forventning til at modparter opretholder de sociale forventninger.

Den fænomenologiske-sociale vinkel tager derimod standpunktet at den sociale relation allerede eksisterer forud for individet og derved forudsat i stedet for at den bliver skabt. I dette synspunkt er tillid noget som opstå mellem individer udenfor deres kontrol og ikke noget som individer på giver hinanden. (Coeckelbergh, 2011, s. 54)

"*The social-phenomenological view I attempt to articulate here, by contrast, defines trust not as something that needs to be 'produced' but that is already there, in the social.*" (Coeckelbergh, 2011, s. 55)

I den kontraktuelle-individualistiske form så er tilliden et resultat af ansvar. I denne form er den skabte tillid grundlagt ud fra et ansvar der er pålagt individet. Et eksempel på den kontraktuelle-individualistiske form kunne være at der er pålagt et ansvar på person B, som kommer af en forventning fra person A, som stoler på at person B vil udføre C, som danner grund for det pålagte ansvar på person B. (Coeckelbergh, 2011, s. 55)

Coeckelbergh (2011) præsenterer gennem teksten de tre ovennævnte tillids-betingelser, som er lavet i forhold til menneske-til-menneske relationer, som skal være til stede for at tillid vil kunne opstå. De tre betingelser er at der skal være en social relation, sprog og friheden til at kunne handle. Betingelsen om sprog er pålagt siden at mennesker gennem sproglig kommunikation kan pålægge og skabe forventninger, løfter og skabe situationer hvor der er påkrævet tillid. Sprog-betingelsen Coeckelbergh omtaler er derfor ikke alment hverdagssprog, men tværtimod en særlig del af sproget hvor at der gennem sproget kan blive pålagt taleren en forpligtelse, som kan fremgå i form af et løfte, som han kalder for moral sprog. I den kontraktuelle-individualistiske form er det nødvendigt med det moralske sprog, da der ellers ikke ville kunne skabes en tillid mellem individer. Den sociale-fænomenologiske form fungerer her som en kontrast til dette, da den både kræver lingvistiske og ikke-lingvistiske forudsætninger for at der vil kunne blive en skabt tillid. (s. 56)

Den følgende betingelse er i forhold til frihed. Friheden skal være specifik i forhold til det individ som der er pålagt et ansvar, da individet skal have frihed til at kunne vælge sine egne handlinger, da der ellers ikke ville være et behov for tillid siden individet så ikke ville have friheden til at vælge at indfri det pålagte ansvar. I tilfældet hvor individet ikke har frihed, så vil der ikke længere være tale om et pålagt ansvar, siden der vil være tale om en ordre og ikke et løfte. Den nævnte frihed på falder også individet der giver tilliden, da dette individ også har friheden til at stole på individet med det pålagte ansvar, siden et individ ikke kan tvinges til at stole på andre. De to individer som indgår i en tillids situation vil derfor ikke kunne kontrolleres eller overvåges i forhold til deres handlinger, siden

dette vil fjerne friheden og derfor også tilliden mellem de to individer. I den kontraktuelle-individualistiske form er specifikt denne betingelse om frihed afgørende. Dette er dog ikke gældende for den social-fænomenologiske form siden den tager et andet udgangspunkt. Her er tilliden og det at individer stoler på hinanden allerede en eksisterende del af den sociale kontekst og det kan derfor ikke kontrolleres af individerne selv. Dette betyder at den usikkerhed som vil befinde sig omkring det at stole på andre ikke er en usikkerhed i forhold til det andet individ, siden det i denne form er en usikkerhed der er relateret til den sociale relation i stedet. (Coeckelbergh, 2011, s. 56)

Den tredje og sidste betingelse omhandler at der er en form for indgået social relation mellem de to individer. Denne betingelse bliver begrundet med at der i en social relation mellem mennesker opstår tillid. I den kontraktuelle-individualistiske form bliver de sociale relationer skabt af de involverede individer, samt at institutioner eller koncepter har samme status, og derfor gælder det også tillid. I den fænomenologiske-sociale form skal der gennem samtaler fremkomme løfter og frihed for det individuelle individ for at der vil kunne være en social relation. Siden der allerede er skabt en social relation, så vil der også forekomme tillid. Sociale relationer eksisterer fordi der er tillid til stede, siden det ikke er tillid der skaber de sociale relationer. (Coeckelbergh, 2011, s. 56)

I den kontraktuelle-individualistiske form fremstår tillid som et valg og det vil derfor ikke kunne være en standard på den måde som det er i den fænomenologiske-sociale form. Dette betyder at der er et valg for individet, siden at den skabte tillid ville kunne brydes, da individet kan revurdere om de forsat stoler på et andet individ, og stoppe denne tillid hvis de føler der er grundlag for at bryde den. Dette vil ikke være tilfældet i den fænomenologiske-sociale form, hvor det her i stedet bliver begrundet at mennesker ikke altid er i fuld kontrol over deres tillid til andre, og at denne mangel på kontrol også gælder for de sociale relationer og hvilken form for relation det er. (Coeckelbergh, 2011, s. 57)

Müller:

"Is it time for robot rights? Moral status in artificial entities" af Vincent C. Müller fra 2021 diskuterer, som titlen antyder, den moralske status i forhold til kunstig intelligens. Han specificerer i starten af sin artikel at diskussionen om rettigheder er særligt vigtigt for at undgå de forfærdeligheder der er foregået igennem historie fra at ske igen. (Müller, 2021, s. 579)

Müller udreder begrebet en fuldbyrdet agent som værende et individ med fri vilje, oftest kaldet en person, som er ansvarlig for sine handlinger. Han nævner yderligere at i diskussionen om robotters rettigheder er den moralske patient af særlig betydning. Forskellen er her at fuldbyrdede agenter har

rettigheder og ansvar hvorimod moralske patienter blot har rettigheder, da skade mod dem har betydning. (Müller, 2021, s. 579)

Derigennem beskriver han at igennem artiklen menes besiddelsen af moralsk status derved både moralske patienter såvel som moralske agenter (fornævnte fuldbyrdet agenter). Det standardiserede synspunkt i forhold til robotters rettigheder er således:

"*If robots had some properties (especially sentience), then they would have moral status, but they do not have those properties now, so they do not have such status now*." (Müller, 2021, s. 580)

Müller beskriver herigennem et argument af Gunkel som foreslår at spørgsmålet ikke omhandler hvorvidt robotterne har rettigheder, men i stedet bør omhandle hvorvidt robotter kan og burde have rettigheder. (Müller, 2021, s. 580)

Herefter påbegynder Müller en rekonstruktion af ovenstående. Han klargør at mennesker har en tendens til at tillægge en form for agenthed, eller i mindste tilfælde værdi, til ting, som for eksempel den måde hvorpå man har særlig respons til sin bamse som barn eller sin yndlingsblyant. (Müller, 2021, s. 581)

Müller opstiller en rekonstruktion af argumentet som lyder således:

1. Jeg føler et ansvar over for X → X har moralsk status
2. Jeg føler et ansvar over for robotter
3. → Robotter har moralsk status

Dette argument er et modus ponens som er et validt argument. Præmis to er desuden i dette tilfælde oftest empirisk sandt. Dog gennem reductio ad absurdum påviser han at argumentet har uacceptable konsekvenser. Dette gør han gennem anvendelse af den tidligere nævnte yndlingsblyant.

1. Jeg føler et ansvar over for X → X har moralsk status
2. Jeg føler et ansvar over for blyanter
3. → Blyanter har moralsk status

Dette argument anvendes for at påpege den måde hvorpå første argument er en form for "Alt er tilladt" argument, som reducerer spørgsmålet om moralsk status til en tilfældig udførelse af vilje, hvor enhver ting som et individ holder en særlig værdi over for, automatisk opnår moralsk status. (Müller, 2021, s. 582)

Herefter gennemgår han en anden teoretikers argument, J. Danaher. Danaher foreslår at den moralske status hverken bør baseres på specifikke kriterier eller mennesker respons på robotterne men nærmere på observerbar præstation. Herigennem fremstilles argumentet:

1. Hvis en robot er nogenlunde performativt ækvivalent til et andet væsen som, gennem generel enighed, har signifikant moralsk status → det er korrekt og passende at tillægge dette væsen rettigheder
2. Robotter kan være nogenlunde performativt ækvivalent til andre væsener som, gennem generel enighed, har signifikant moralsk status
3. → Det kan både være rigtigt og passende at tillægge robotter signifikant moralsk status

Dette beskriver Müller som værende et lignende argument til "hvis det går som en and, svømmer som en and og kvækker som en and, så har den moralsk status som en and". (Müller, 2021, s. 582)

Müller uddyber yderligere at blot at have tilstrækkelig grund til at tro at det er en and ikke er nok til at tillægge moralsk patienthed til væsnet. Særligt for robotter nævnes også at disse endda kan være designet specifikt til formålet om at overbevise mennesker om at den har moralsk status. Derudover nævnes at dette argument fremstiller moralsk status som noget mennesker kan tillægge andre væsener og ikke blot noget som robotten enten besidder eller ikke besidder, hvor der stilles spørgsmålstegn til hvorvidt mennesker besidder sådan en kraft af Müller. (Müller, 2021, s. 583)

Müller påpeger desuden et problem som denne opgave også vil omhandle, at der endnu ikke findes en videnskabelig metode som kan teste for fænomenal bevidsthed eller fri vilje, og at dette har en betydning for enhver diskussion af tillæggelse af moralsk status. Han nævner herigennem at hverken det tidligere nævnte "alt er tilladt" argument eller et bredt forebyggende princip ville være en mulighed. Han foreslår i stedet at der, på samme måde som med objekter af særlig betydning, kunne opstå hensyn. Opførsel over for sådanne objekter kan stadig dømmes som moralsk rigtigt eller forkert, men at den moralske status af betydning her i stedet placeres på mennesket som har moralsk status og som opfatter robotten som besidder af værdi. (Müller, 2021, s. 584)

Schneider:

Den første del af bogen "Ethics of Artificial intelligence" fra 2020 som denne opgave vil arbejde med er "How to Catch an AI Zombie: Testing for Consciousness in Machines" af Susan Schneider. Dette kapitel omhandler testning af bevidsthed i kunstig intelligens.

I kapitlet definerer Schneider begyndelsesvis hvilken form for bevidsthed som kapitlet omtaler. Denne bevidsthed er en form for bevidstheds oplevelse eller følt bevidsthed. Det er derfor hverken påkrævet at vente på en formel definition af bevidsthed eller opnå en komplet forståelse af den neurale basis for bevidsthed. Det defineres som dette på den begrundelse at intet af ovenstående er påkrævet for at udpege bevidsthed i maskiner. (Schneider, 2020, s. 439)

Schneider påpeger vigtigheden af at uddybe forståelsen af maskiners bevidsthed og beskriver herigennem besværlighederne der opstår i et forsøg på det. Her nævnes det, i denne opgave, tidligere nævnte Black Box Problem der opstår i deep-learning kunstig intelligens. Hvis det var muligt at fremstille den kunstige intelligens' kognitive og perceptuelle arkitektur, ville mennesker stadig have besværlighed med at identificerer funktioner som er centrale for bevidsthed. Selv med kognitive funktioner som opfattes som vigtige for biologisk bevidsthed som eksempelvis opmærksomhed og fungerende hukommelse, ville det ikke nødvendigvis være mere end blot en antydning på at robotten er bevidst. (Schneider, 2020, s. 442)

Schneider uddyber yderligere i kapitlet forståelsen af følt bevidsthed igennem den først beskrevne test, ud af tre. Den følte bevidsthed er defineret som den måde det føltes internt i et menneske at opleve verdenen. Den måde hvorpå mennesker kan forestille sig andre oplevelser og den indlevelsesevne som mennesker besidder har betydning for denne følte bevidsthed. (Schneider, 2020, s. 442)

Den først nævnte test som vil forsøge at teste efter denne følte bevidsthed, er ACT. Denne test beskrives som en spørgsmål-svar test, hvor adskillige spørgsmål vil blive spurgt. Ethvert tilfredsstillende spørgsmål ville være tilstrækkeligt for at bestå testen. Et eksempel på spørgsmålene kunne være: om maskinen opfatter sig selv som andet end blot dets fysiske selv, spørgsmål om reinkarnation eller ud-af-kroppen-oplevelser. (Schneider, 2020, s. 443)

For at opnå et acceptabelt resultat ville det være nødvendigt at fjerne den kunstige intelligens' adgang til internet og andre former for dataindsamlings muligheder. Det er dog nødvendigt at den kunstige intelligens har naturligt sprog som den kan kommunikere dens indre tilstand gennem. (Schneider, 2020, s. 444-445)

Schneider argumenterer for at ACT-testen kan være nyttig under udviklingen af forskellige former for kunstig intelligens, både for at undgå at bruge bevidste robotter på en uetisk måde, men også for at assistere i at skabe kunstig bevidsthed. Derudover nævnes også en anden udgave af denne test som

tester en større gruppe af kunstig intelligens. Denne test er nærmere til sidst nævnte formål da der her bliver lagt fokus på hvordan disse kunstige intelligenser sammen ville udvikle sig gennem interaktioner med hinanden. (Schneider, 2020, s. 446)

Begrænsningerne for denne test er særligt fokuseret omkring kravet om sprog, og at en mangel på dette ikke nødvendigvis afviser at robotten stadig kunne opleve følt bevidsthed, på samme måde som børn og andre ikke-menneskelige dyrearter ikke kan kommunikere deres følte bevidsthed. Derudover nævnes også at ACT er baseret ud fra et menneskeligt synspunkt på bevidsthed og at spørgsmål og situationer som robotten ville blive udsat for, er baseret ud fra menneskers opfattelse af fænomenal bevidsthed. (Schneider, 2020, s. 447)

Schneider beskriver herigennem et vigtigt skel mellem den ovenstående fænomenale bevidsthed og et begreb indenfor maskiners bevidsthed, kognitiv bevidsthed. Den fænomenale bevidsthed er som tidligere beskrevet den måde det føles indeni at være dig, hvorimod kognitiv bevidsthed er en beskrivelse af en kunstig intelligens som besidder de arkitektoniske funktioner der understøtter menneskers fænomenale bevidsthed. Disse funktioner er, som tidligere nævnt, opmærksomhed og fungerende hukommelse. Dette skel er vigtigt, da det kan argumenteres at hvis den kunstige intelligens besidder kognitiv bevidsthed, så er der et potentiale for at den også besidder den fænomenale bevidsthed og at det her ville være vigtigt at teste. (Schneider, 2020, s. 448)

Den anden nævnte test er IIT (Integrated information test), som omhandler niveauet af gensidig afhængighed af systemets dele, med et højt niveau af feedback imellem delene. Niveauet bliver målt (og beskrevet gennem det græske tegn Φ) og testen skulle herigennem kunne påvise både bevidsthed samt niveauet af bevidsthed. Problematikken med denne test er at niveauet af bevidsthed ikke kan bedømmes præcis medmindre systemet er ekstremt simpelt. Schneider påpeger dog at det, på trods af denne begrænsning, ikke nødvendigvis er en betydningsløs test, da der stadig kunne være et påkrævet minimums niveau af Φ for at noget kunne defineres som værende bevidst. (Schneider, 2020, s. 449-450)

Den tredje og sidst nævnte test er Chip testen. Denne test er baseret i den menneskelige bevidsthed, og er tættere knyttet til udviklingen af bevidste robotter. Denne test involverer at erstatte dele af den menneskelige hjerne med neurale proteser (i form af chips) for at undersøge chippens mulighed for at kunne understøtte fænomenal bevidsthed. Hvis en sådan chip findes, ville det betyde at enhver kunstig intelligens med samme chip indbygget bør testes nærmere for fænomenal bevidsthed, for eksempel gennem ACT eller lignende test. (Schneider, 2020, s. 452)

Schneider nævner ud over disse tests at et forsigtighedsprincip bør indføres. Dette princip påkræver at hvis der er en chance for at en teknologi kan forsage katastrofisk skade, så er det påkrævet at udviklerne af en sådan teknologi beviser at deres produkt ikke vil have denne effekt. Schneider foreslår herigennem at der både testes for bevidsthed såvel som effekten af en sådan bevidsthed i forbindelse med empati og pålidelighed. Denne testning for bevidsthed bør være fortsættende. Det er vigtigt for at undgå både uetiske handlinger overfor robotterne såvel som undgåelse af eksistentielle risiko for menneskeheden. Derudover foreslås at testning bør være en almindelig del af udvikling og research indenfor sofistikerede kunstig intelligens systemer, samt at der lægges fokus på sikkerhed i disse test. Hvis en kunstig intelligens opdages som værende bevidst så bør denne herigennem opnå samme rettigheder som andre følende væsener. Sidst men ikke mindst, hvis der opstår tvivl om hvorvidt en kunstig intelligens er bevidst, bør samme rettigheder påfalde den kunstige intelligens. (Schneider, 2020, s. 454-455)

Liao:

"Ethics of artificial intelligence" af S. Matthew Liao fra 2020 handler om nogle af de etiske problemstillinger som kan opstå i forbindelse med kunstig intelligens. I denne opgave vil der kun blive arbejdet med den fjerde del af bogen, som specifikt er kapitel 17: "The Moral Status and Rights of Artificial Intelligence". (og kapitel 15 i form af Schneider) Dette kapitel kommer ind på nogle af de rettigheder der potentielt kunne komme i fremtiden i forhold til robotter og hvordan man kan gå ind og begrunde sådanne rettigheder. Liao's begrundelser er baseret på et art neutralt krav, hvilket betyder at det at være menneske ikke danner et grundlag for en ret til rettigheder. (s. 482)

Liao (2020) begrunder hvordan det at være berettiget til rettigheder ikke skal være fokuseret på intrinsiske værdier, så der ikke er et medfødt krav til rettigheder. Liao fokuserer i bogen på hvordan den kunstige intelligens bliver udviklet, og hvordan udviklingen der har foregået de seneste år på sigt vil kunne gå en fremtid i møde, hvor der ville være behov for retskrav. Hvis der på sigt ville blive udviklet en kunstig intelligens som opnår moralsk agenthed, og som derved ville ligne mennesker, så ville den kunstige intelligens opfylde kravet til at have rettigheder, og det ville derved være nødvendigt at have rettigheder baseret på intrinsiske værdier. (s. 482)

"Moral status is the standing an entity has that gives moral agents a pro tanto reason to act toward it in a certain way." (Liao, 2020, s. 480)

Moralsk status ville skabe en stor ændring i det nuværende samfund, da det ville skabe en ændring i hvordan vi agere overfor kunstig intelligens. Liao (2020) fremlægger det som, at den moralske status

der er pålagt mennesker, danner grundlaget for den måde mennesker agere mod andre dyr af lavere moralsk status, og hvordan mennesker behandler andre mennesker, der besidder samme moralske status. Han beskriver herigennem også hvordan den moralske status har betydning for betydning for retten til rettigheder og mængden af dem. Dette kommer for eksempel til udtryk gennem hunde, hvor at hunden besidder lavere moralsk status end mennesker, men det betyder ikke at mennesker kan forårsage smerte på hunden mod egen nydelse, hvilket er på grund af hundens moralske status. (s. 480-481)

Liao (2020) forslår gennem dette kapitel nogle aktuelle intrinsiske egenskaber, der ville kunne danne grundlag for at få retskrav. Liao begrunder 6 forskellige intrinsiske egenskaber, som kunne danne grundlaget for retskrav: at være levende, have bevidsthed, at kunne føle smerte, at kunne have egne ønsker, være i stand til moralsk agenthed og at være i stand til rationel agenthed. (s. 482-483)

Liao (2020) ser sine forslag som en genetisk basis til moralsk agenthed som betingelser for retskrav. Dette forslag kommer udenom nogle af de problematikker som andre filosoffer får i forbindelse med ekskludering af nogle bestemte menneskegrupper, hvor eksempler på dette kunne være komapatienter, spædbørn og mentalt handicappede. Speciasistiske betingelser bliver samtidig også undgået, hvor et eksempel kunne være en betingelse der ikke kræver andet end at individet er et menneske. Hvor Liaos betingelse for individer er at de skal have genetisk basis for moralsk agenthed. (s. 484)

I Liaos optik, så er den moralske agenthed som mennesker besidder fundet i deres genom:

"In human beings, this set of codes [physical codes that generate moral agency] is located in their genome. We know this because the developmental basis for adaptive phenotypes like moral agency requires a great deal of complexity, and the genome contains a significant proportion of this complexity." (Liao, 2020, s. 484)

Siden at alle mennesker besidder det omtalte genom, så betyder det at mennesker har en genetisk basis for moralsk agenthed. Der vil derfor ikke være muligt at kunne udelukke menneskegrupper som komapatienter, spædbørn og mentalt handicappede, da de vil besidde retten til retskrav, siden de er i besiddelse af det omtalte genom. Det betyder også samtidigt at andre arter ikke vil kunne blive ekskluderet ud fra begrundelsen at af de tilhøre en anden art. Liao fremstiller i denne forbindelse, at hvis man i fremtiden gjorde opdagelsen at en gorilla var i besiddelse af moralsk agenthed, så ville det betyde at gorillaen ville opnå den samme genetiske basis som mennesker, og det samme gældende

for eventuelle rumvæsner. Dette betyder derfor at den genetiske basis ville kunne blive anvendt til at undersøge om en kunstig intelligens besidder genetisk basis for retskrav. (Liao, 2020, s. 484 – 487)

Liao (2020) fremstiller tre forskellige måder hvor igennem at kunstig intelligens ville kunne opnå den genetiske basis for moralsk agenthed. De tre metoder er: bevidsthedsuploading, gradvis udskiftning og at programmere en moralsk kunstig intelligens. (s. 486)

Indenfor udvikling af kunstig intelligens og generel teknologi har der være en gradvis voksende interesse indenfor bevidsthedsuploading. Dette opnås gennem at skabe en præcis detaljeret kopi af et individs hjerne. På den måde som den gældende hjerne fungerer, overføres til software og bliver uploadet til en simuleret verden, eller alternativt uploadet til en robot, som derefter kan interagere med verden omkring den. I tilfældet hvor den menneskelige hjerne er blevet fuldstændig kopieret, så ville der være tale om en robot med genetisk basis for moralsk agenthed, da det er en fuldkommen kopi af et menneske som bliver imiteret. (Liao, 2020, s. 487 – 488)

Den anden metode gradvis udskiftning adskiller sig fra bevidsthedsuploading, da der her er tale om en gradvis erstatning af hjernen, hvor der i den første metode var tale om en kopiering. Denne metoder arbejder frem mod en erstatning af hjernen over tid, hvor de kulstofbaserede celler vil blive erstattet af uorganiske celler over tid, indtil hjernen kun består af uorganiske celler. Hvis der gennem denne proces ikke sker nogen ændringer til bevidstheden, samt livsvigtige funktioner som stofskifte og optagelse bliver bevare, så ville der også være tale om en genetisk basis for moralsk agenthed, siden den bliver bevaret fra før den gradvise erstatning startede. (Liao, 2020, s. 488)

Den sidste metode handler om at skabe en kunstig intelligens gennem programmering, som har moralsk agenthed, og derfor skaber grundlag for genetisk basis. Liao (2020) kommer herigennem med to forskellige metoder, som potentielt ville kunne resultere i kunstig intelligens med moralsk agenthed. Den første metode ville være at programmere en kunstig intelligens, der indeholder alle de etiske teorier, som for eksempel deontologi og konserkventialisme. Dette ville potentielt kunne skabe en basis for moralsk agenthed, men samtidig pointerer Liao at der ikke er konsensus omkring hvilke etiske teorier der er korrekte, hvilket ville betyde at der ikke kan skabes moralsk agenthed ud fra disse teorier for en kunstig intelligens. Den anden foreslåede metode er at lære en kunstig intelligens moralsk agenthed ved brug af samme fremgangsmåde som sprog læres. Den kunstige intelligens ville med denne metode ikke lære omkring etiske teorier, men den ville i stedet lærer omkring hvad der antages at være moralsk rigtigt og forkert og derigennem lære en form for moralsk grammatik. Den kunstige intelligens ville derfor lære omkring handlinger med forskellige variabler som agent, tro,

intention, handling, modtager, konsekvens og moralsk evaluering. Disse variabler vil blive komprimeret til en mekanisme lignende moralsk bedømmelse, som vil resultere i en af tre udfald: tilladeligt, utilladeligt og påkrævet. (s. 488 - 489)

Liao (2020) udvælger i kapitlet nogle rettigheder som en kunstig intelligens ville kunne opnå og som den potentielt ville kunne blive påkrævet. Her i blandt er mange af de nævnte rettigheder tilhørende de menneskelige rettigheder, som også ville kunne inkludere kunstig intelligens, men også rettigheder der ville være målrettet specifikt mod kunstig intelligens. Liao pointerer her at der ikke ville være tale om juridiske rettigheder men menneskerettigheder. (s. 491)

Menneskerettigheder danner grundlaget for de fundamentale betingelser, som danner rammerne for et godt liv. Liao (2020) definerer et godt liv som værende grundlæggende aktiviteter som anses som at være værdifulde. Eksempler på disse aktiviteter kan være dybe personlige forhold, viden, aktiv nydelse som eksempelvis kreativt arbejde og passiv nydelse som kunne være at sætte pris på skønhed. For at kunne opnå disse ovenstående fundamentale kapaciteter, så er det påkrævet at kunne: kapacitet til at tænke, at kunne være motiveret af fakta, at kunne vide, at kunne vælge og agere frit, at sætte pris på værdien af noget, at kunne udvikle interpersonelle forhold og at have autonomi. Disse fundamentale kapaciteter vil derfor også kunne bruge til at forklare hvorfor menneskerettigheder skal ses som valide og bør følges. I tilfældet hvor en kunstig intelligens kan opfylde disse kapaciteter, så ville det derfor også være nødvendigt at have lignende rettigheder for kunstig intelligens. Rettighederne for en kunstig intelligens vil ikke nødvendigvis være de sammen som dem for mennesker, da alle ikke nødvendigvis ville have betydning for en kunstig intelligens, men derfor ville flere af dem stadig kunne være brugbare. (s. 491 - 495)

Derudover udforsker Liao (2020) også scenarioet hvor en kunstig intelligens opnår en højere moralsk status end mennesker. Denne ide kommer fra konceptet at en kunstig intelligens ville kunne opnå viden omkring deres skabelse, og derved potentielt ville kunne skabe ny kunstig intelligens mere et højere niveau af moralsk status, siden den kunstige intelligens ville være skabt med menneskeligt niveau af moralsk status. (s. 495 - 497)

Seibt, Damholdt & Vestergaard:

I teksten "Integrative social robotics, value-driven design and transdiciplinarity" af Johanna Seibt, Malene Flensborg Damholdt og Christina Vestergaard præsenterer de konceptet af Integrative Social Robotics (ISR). ISR er en kontrollerings og regulerings måde at udvikle sociale robotter på, som indeholder nogle principper der burde følges hvis man som individ arbejder indenfor research, design

og development (RRD) af robotter. Metoden er lavet ud fra et non-replacement principle: ”Social robots may only do what humans should but cannot do” (Seibt et al., 2020, s. 111) Princippet indebære at udvikler robotter til at udføre opgaver, som burde blive gjort af mennesker, men som mennesker ikke udfører, og til gengæld burde robotter ikke fjerne opgaver eller overtage sociale funktioner fra mennesker. Dette vil resultere i robotter som en forlængelse af sociale funktioner og ikke en erstatning af mennesker. Et eksempel på dette kan være robotter brugt til genoptræning, hvor præcise gentagelser kan være nødvendige. (Seibt et al., 2020, s. 111, 131 - 132)

Formålet med ISR er derfor mere en risiko minimering end andre tiltag er. Dette kommer til udtryk i form af at ISR researcher hvad robotterne burde kunne gøre i samfundet, hvor den nuværende research mere omhandler hvad robotter er i stand til at gøre. Det betyder at i stedet for at udviklere af robotter ser på robotten efter udvikling, så vil ISR skabe et system robot-udviklere følger under udviklingen af robotter. Denne måde at arbejde på som robot-udvikler burde fjerne mulighederne for at skabe en robot som kunne udgøre en fare eller have en negativ påvirkning på samfundet. (Seibt et al., 2020, s. 114 - 115)

ISR bygger på fem principper; Processprincippet, kvalitetsprincippet, ontologisk kompleksitetsprincippet, kontekstprincippet og værdier først princippet. Det første princip, process princippet (P1), handler om fokusset under udviklingen, hvor det bør være på sociale interaktion og ikke på robotten. Det vil sige at robotten selv ikke er i fokus, dette er siden der er stadig ikke er nok information omkring sociale robotter, samt hvordan de virker. Fokusset vil i stedet være på hvilke interaktioner robotten er en del af og hvordan robotten agerer i disse situationer. (Seibt et al., 2020, s. 126)

Kvalitetsprincippet (P2) skaber et krav om ekspertviden gennem RDD, hvilket betyder at alle relevante discipliner robotten agerer indenfor. Dette betyder at hvis man vil udvikle en robot, som ville skulle bruges indenfor plejesektoren, så vil det være et krav at der under udviklingen bliver brugt expertise indenfor pleje. (Seibt et al., 2020, s. 127 – 128)

Ontologisk kompleksitetsprincippet (P3) bygger på at alle sociale interaktioner indebærer mindst tre individer, hvor to interagerer med hinanden og en tredje person fungerer som en ekstern observatør. Derfor bør RDD processen inkluderer at de sociale interaktioner robotten foretager er skabt efter dette princip. (Seibt et al., 2020, s. 129 - 130)

Kontekstprincippet (P4) handler omkring identiteten af en social relation og hvordan den specifikke identitet er relativ til konteksten. Det vil derfor være nødvendigt at man foretager regelmæssig feedback og tilrettelser i forhold til de sociale robotter. (Seibt et al., 2020, s. 130 – 131)

Værdier først princippet (P5) omhandler som beskrevet tidligere, at når robotter bliver udviklet så må det ikke være med formålet at erstatte mennesker, men i stedet bør de fungere som en forlængelse af menneskets arbejde og anvendes i bestemte situationer, hvor mennesker ikke bør handle. Dette non-replacement requirement skal bruges af udviklere, som arbejder med sociale robotter. (Seibt et al., 2020, s. 131)

Seibt, Damholdt og Vestergaard (2020) beskriver derudover også at målet og det som prøves at blive opnået med kunstig intelligens ikke kan ses som realisme, men bør ses som at robotterne bliver udviklet til så højt et niveau, at de vil kunne imitere et så højt niveau af de ville kunne snyde mennesker. I produktionen af sociale robotter, så sigter man efter at imitere menneskelighed eller at efterligne de menneskelige kvaliteter. (s. 121)

I teksten er der blevet udarbejdet fem spørgsmål som kan bruges til at evaluere, så man derved kan identificere potentielle fare ved skabelsen af sociale robotter. Disse spørgsmål er:

1. *How do interactions with social robots affect individual and social human well-being, including the ethical and existential dimensions of such well-being?*
2. *How will interactions with social robots affect human-human interactions […], will we have less empathy or will human empathy no longer rely on direct perception?*
3. *How will interactions with social robots affect our cultural conceptions of a social interaction, e.g., will the notion of sincerity become obsolete?*
4. *How will interaction with social robots affect our value system, e.g., will we tie human dignity more closely to autonomy, which social robotics can enhance, instead of keeping it linked to interpersonal recognition?*
5. *How will our interactions with social robots affect our emotional, cognitive, and physical capacities, differentiated for different developmental phases of a humanlife?* (Seibt et al., 2020, s. 123)

Disse spørgsmål fremstår som en barrikade for muligheden for at skabe lovgivende handlinger, da der spørgsmålene ikke kan besvares på grund af manglende research, og der vil ikke kunne laves nogen lovgivning før man kan besvare disse spørgsmål. Dette skaber dog samtidigt et problem i

forhold til udviklingen af sociale robotter, da det ikke vil være muligt at kunne forudse de potentielle farer som ville kunne forekomme under udviklingsprocessen.

Diskussion:

Coeckelbergh:

En af de betingelser som Coeckelbergh nævner i hans artikel ”Can We Trust Robots?” er sprog. Det er, som tidligere nævnt, ikke blot sproget som praktisk evne, men i stedet en form for moralsk sprog som mennesker anvender igennem tillidsformning. Mennesker skaber igennem deres ord tillidsbygning i form af løfter og lignende. Et eksempel på dette kunne være ”Du kan stole på mig, jeg lover at jeg vil gøre X”. Denne sætning er et eksempel på den type af moralsk sprog som Coeckelbergh henviser til i hans artikel.

Det moralske sprog, som lingvistisk koncept, ville ikke umiddelbart være uopnåeligt for humanoide robotter. Tankerne bag denne type er sprog er dog sværere at opnå og argumenterer for foregår. Mellem mennesker er der oftest en forståelse af at ordene i disse udtalelser ikke blot er ord, men at de medfører en handling som er blevet lover på forhånd og som bør udført, hvis man indfrier sine løfter. Det moralske sprog er derfor både indeholdende en forventning såvel som løftet. Forventningen påfalder individet som bliver lovet noget og dette fungerer som en form for incitament til at indfri løftet af løfte-skaberen. Det er en særligt menneskelig viden at denne forventning kan udmunde i skuffelse, skulle løftet ikke indfries, og denne viden skaber, for løfte-skaberen, en form for følelse. I det menneskelige findes der det medmenneskelige og dette har en påvirkning på vores forståelse af moralsk sprog. Her findes betydningen for dette koncept ikke blot i ordene men snarere i menneskerne som udtaler ordene.

Humanoide robotters evne til at kunne opnå dette niveau af følelsesregister for andre individer er et koncept som, på trods af besværlighed, stadig er højst nødvendigt at forsøge at finde et svar til. En følelse som er essentiel i denne sammenhæng, såvel som andre, er empati. I denne opgave defineres dette som en form for medfølelse som mennesker oplever overfor hinanden. Det er netop denne menneskelige egenskab som kan skabe en forskel mellem et menneske og en robot, i forbindelse med moralsk sprog. Selvom robotten, teoretisk set, kan udtale ordene ”Jeg lover dig X” og efterfølgende indfri dette løfte, betyder det ikke at forståelsen af moralsk sprog er opnået. En sådan forståelse påkræver et niveau af bevidsthed som ikke begrænser sig til en generel selvforståelse. Forståelsen kræver i stedet en form for social forståelse, der overskrider en almindelig opfattelse og oplevelse af en social interaktion. Mennesker er ikke blot individer, men sociale individer der

indgår i en særligt menneskelig sammenhæng. Ikke-menneskelige dyr har hver især en udgave af en sådan sammenhæng men den menneskelige socialitet udskiller sig ved at være mere end blot samvær og hierarki. Selvom medfølelse kunne argumenteres for at have fluktueret i niveau igennem menneskers historie, er det stadig til stede i højere eller lavere niveauer uanset tidspunkt eller årstal.

Humanoide robotter ville herfor både skulle kunne opnå en sådan empati-følelse fra et andet individ, robot eller menneske, og samtidigt også selv opnå samme empati. Det ville ikke være uopnåeligt for robotten at opnå empati og medfølelse fra andre mennesker, da medfølelse er en følelse som mennesker oftest tillægger både andre mennesker og andre væsener. Her kunne nævnes det særlige bånd mennesker har for deres kæledyr, som en del af deres hverdag og den måde hvorpå selv andre dyr, som afskåret fra individet, stadig inspirere medfølelse.

Den empati som mennesker oplever overfor andre mennesker er, på trods af dette, sværere at opnå. Det er den af den årsag at den involverer en indlevelse som er anderledes end den forvrængede udgave som sker i antropomorfiske situationer med dyr. Den humanoide robot ville i dette tilfælde være påkrævet at være et følende væsen, både i form af nydelse og smerte, og yderligere i form af det emotionelle spektrum som transcenderer disse. Et spektrum som ville være påkrævet ville involvere følelser som tristhed, nervøsitet, glæde og empati for at der kunne argumenteres for at de humanoide robotter har opnået fænomenalt bevidsthedsniveau, som er det menneskelige. Disse følelser kan have stor betydning for hvorvidt de humanoide robotter ville opnå forståelse for moralsk sprog, som det forstås i den bredeste forstand.

Coeckelberghs anden betingelse for tillid er frihed. Dette er et diskussionsfelt som har særlig betydning for humanoide robotters rettigheder. Fuld autonomi er et koncept indenfor kunstig intelligens som skaber frygt i mange mennesker. Frygten er veldokumenteret i henholdsvis bøger, computerspil og film og har haft stor fokus i diskussioner omkring humanoide robotter, selv før konceptet synes opnåeligt i nogen som helst udgave. Et kendt og særligt afbenyttet koncept indenfor science fiction er humanoide robotter som beslutter sig for at overtage verden eller i den mindste grad overtage styringen over mennesker. Det er forsat et koncept som bliver benyttet af genren.

Coeckelbergh beskriver frihed som muligheden for at handle efter egen vilje, udenfor andres overvågning og kontrol. En frihed som denne ville påkræve at de humanoide robotter blev skabt med fuld autonomi, og tilladt mulighed for at reagere uden intervention. Det er dog blot nødvendigt, for Coeckelbergh, med quasi frihed for at skabe basis for tillidsbygning. Han beskriver quasi frihed som værende tilstrækkeligt at begge individer i den sociale interaktion oplever hinanden som frie.

Det er derfor ikke en nødvendighed at begge individer faktuelt har frihed. Quasi frihed er et koncept som tilskrives en problematisk tankegang i forbindelse med en fremtid hvor robotterne opnår bevidsthed. Her er manglen på interesse for fuld autonomi for robotterne en tilstand som kunne føre til en restriktion på deres oplevelse af deres menneskelige moralske status, og som begrænser deres rettigheder i forbindelse med retten til at være fri.

En yderst vigtig del af de tre betingelser er den sociale relation. Det er den på den baggrund at hvis denne ikke er til stede, frafalder behovet for skabelsen af tillid og behovet for tillid totalt. Årsagen bag dette er at hvis mennesker ikke opfatter robotten som et individ og i stedet nærmere som en ting, så frafalder den sociale relation til robotten. Det er dog samtidigt vigtigt at pointere at denne betingelse ikke nødvendigvis er svær at opnå for robotterne. Menneskers indlevelsesevne er massiv og en indlevelse i robotterne som individer ville hurtigt kunne udvikle sig til en realitet. Det er en stor fortaler for denne udvikling at mennesker i forvejen har tendens til at menneskeliggøre objekter og ikke-menneskelige dyr og antropomorfisme ville i dette tilfælde være en positiv udvikling. En antropomorfisk tilgang kunne assistere med at danne den empati som blev nævnt tidligere og skabe fællesskabsfølelse mellem de humanoide robotter og menneskerne som indgår i en social relation med dem.

Hvis de tre ovenstående betingelser bliver opfyldt, vil det hentyde til at der er skabt en individ-til-individ relation og at tillid kan opstå mellem arterne. Det kan herigennem argumenteres at skulle denne tillid opstå, så ville der allerede have været skabt en menneske-til-menneskelignende association mellem de humanoide robotter og menneskerne. Denne menneske-til-menneske relation er af yderst stor betydning for robotternes fælles oplevelse med mennesker, såvel som kvaliteten af det samvær som opstår.

Coeckelbergh nævner yderligere i hans artikel den kontraktuelle-individualistiske form. I denne form dannes ansvaret som et resultat af tillid. Det er derfor, som navnet antyder, individer som indgår i en form for kontrakt med hinanden gennem tillidsbygning hvor kontrakten pålægger løfte-skaberen ansvaret for den sociale interaktion. I den kontraktuelle-individualistiske form opstår der ingen tydelige problemer for humanoide robotter, med betingelsen at robotterne bliver opfattet som individer. Det kunne i stedet potentielt tilvejebringe at konceptet er letter for den humanoide robot at begribe, selv i tilfælde hvor robotten mangler evnen til at fortolke implicit nonverbal kommunikation.

I den fænomenologisk-sociale form er det fornævnte ansvar i stedet allerede indbygget i det sociale. Ansvaret er her snarere en implicit nonverbal kommunikeret tilstand som kunne forsage komplikationer for den humanoide robot, i forbindelse med forståelse. Begrundelsen for at sådanne komplikationer kunne opstå er at nonverbal kommunikation er en variabel som kan ændre sig alt efter kultur og individ. Det nonverbale sprog er en særlig kategori af sprog som ville være sværere at definere og generalisere. Nonverbale tegn på tillid kan variere i forskellige grader fra individ til individ og dette kunne skabe miskommunikation hvis robotten ikke fortolker disse variationer korrekt. Det bør herigennem nævnes at nonverbal kommunikation eller kropssprog beskriver et felt som ikke altid fortolkes korrekt mellem mennesker og at det potentielt også kunne være tilfældet at robotten ville absorbere og forstå denne form for kommunikation bedre end mennesker. Årsagen til denne argumentation er at nonverbal kommunikation er mønsterpræget og at dataen ville kunne gemmes i robotten på en måde som ikke er muligt i mennesker. Det er samtidigt også en informationskilde som oftest foregår forholdsvis ubevidst i mange sociale interaktioner mellem mennesker og det ville umiddelbart være påkrævet at robotten kunne få det observerede mønster beskrevet og bekræftet da det ellers ville forblive ubeskriveligt og essentielt set ubrugeligt.

De neurale netværker, som blev beskrevet i redegørelsen af denne opgave, ville i dette tilfælde indoptage data i forbindelse med samværet med et andet individ og derefter kunne analysere denne data og producere et output. Et output, i denne sammenhæng, kunne for eksempel være et bestemt nonverbalt udtryk som blev observeret som et mønster af den kunstige intelligens. Det uklare ville i dette tilfælde være hvordan mønstret skal kunne bekræftes og om dette ville være en mulighed for robotten selv at bekræfte. Hvis dette dog formodes muligt, ville den intense indsamling af data kunne føre til eksponentiel udvikling indenfor forståelsen af nonverbal kommunikation, som både kunne berige menneskelig forståelse såvel som andre humanoide robotter. Dette ville tildele begge yderligere viden om denne type af kommunikation og de betydninger som findes i adfærden, især hvis denne data kunne blive gjort tilgængelig for nonverbale kommunikationsforskere.

Müller:

Müller fremlægger Danahers argument for moralsk status i kunstig intelligens og sammenligner dette med det velkendte ordsprog ”Hvis det går som en and, svømmer som en and og kvækker som en and”. Müller fremsætter at blot fordi det går som en and, svømmer som en and og kvækker som en and, betyder det ikke at væsnet er en and. Der er dog en mangel på begrundelse for hvorfor mennesker ikke er tilladt at behandle væsnet som en and. Dette er af særlig betydning fordi anden såvel som den kunstige intelligens i hans eget eksempel blot har moralsk status som moralsk

patient, som han også selv pointerer. Moralsk patienthed, som forklaret tidligere i opgaven, påkræver ikke hverken rettigheder eller andre krav af stor betydning. Det påkræver blot af mennesker at væsenet behandles med respekt og ikke misbruges. Danaher argumenterer for den samme signifikante moralske status som mennesker har, hvor den kunstige intelligens ville være sin egen agent med fulde rettigheder. Grundet Müllers egne begrænsning til en and, vil denne del af opgaven derfor i stedet fokusere på den moralske patienthed.

Problematikken i ovenstående ordsprog, i forhold til kunstig intelligens, er at der ikke blot er tale om noget der opfører sig som en and. Gennem opgaven er der redegjort for hvor besværligt og potentielt fejlslået testningen for bevidsthed kan være og der kan derfor argumenteres for at væsenet i virkeligheden er en and, men at mennesker blot ikke har kunne bevise det. For at anvende Müllers egen beskrivelse, så kan der argumenteres for at vi blot ikke har fundet den type af and før og derfor føler os nødsaget til at betegne det som noget andet end en and. I tilfældet af kunstig intelligens ville sådan en fejl have katastrofale uetiske konsekvenser.

Det selvmodsigende i hans argument er den, tidligere i artiklen, bemærkning om historie. Her nævnes hvordan der i gennem historien har været tilfælde hvor uenigheder omkring moralsk status af diverse individer og væsener har ført til grotesk behandling af mennesker, såvel som dyr. Det er ikke nødvendigvis selvmodsigende hvis der er tale om supervised læring og testning under udvikling hvorefter robotten forbliver i en standardiseret udgave. Dette er absolut profitabelt hvis robotten blot fungerer som en arbejder, men er der tale om en social robot problemet langt mere komplekst. Som nævnt tidligere påkræver supervised læring at hvert et billede, lyd eller ord bliver defineret gennem annotering. Dette ville umiddelbart være en massiv opgave når robotten skal fungere som en social robot. Det vil derfor være langt mere rentabelt at udforme robottens standardiserede udgave gennem unsupervised læring. Samtidigt ville det være vigtigt med sociale robotter at de udvikler sig ud over den standardiserede udgave for at tilegne sig sin respektive opgave. Dette kunne for eksempel være en personlig plejerobot. Her ville det være vigtigt for kvaliteten af plejen at robotten lærer af sine erfaringer med patienten og herigennem udvikler sig ud over den standardiserede udgave. En sådan udvikling ville også foregå gennem unsupervised læring og her påbegynder problematikken med en potentiel spontan udvikling af bevidsthed. Argumentet for at en sådan udvikling kunne ske spontant er at grundet den manglende viden om bevidsthed i sin helhed, ville det være umuligt at definitivt afvise at bevidsthed ikke kunne udvikles over tid.

Det ville af denne årsag være langt vigtigere at sikre mod potentielle katastrofale konsekvenser, end det er om robotten i sig selv faktuelt har moralsk status af nogen art. Særligt hvis der er tale om moralsk patienthed, hvor det blot er påkrævet at robotten behandles med respekt og at skade mod den anses som moralsk forkert. Her nævner Müller selv hensyn som et alternativ. Problematikken med dette er bestående i dets afhængighed af mennesket. Det ville betyde at den eneste det afskrækker en bruger af robotten fra at udføre skade mod den, er brugerens egen selvrespekt for sine ejendele. Ejerskab over en social robot med fænomenal bevidsthed er i sig selv problematisk, og bestemt en problematik som er sværere at undgå, selv gennem moralsk patienthed da det er tilladt at eje arter som befinder sig i denne moralske status. Det kan dog tilskrives som en sikring af ikke-bevidste robotter og derved blot en midlertidig begrænsning for robotter som udviklede bevidsthed.

Müller nævner hvorvidt mennesker har kraften i forhold til at give andre væsener moralsk status og den almindelige konsensus er umiddelbart at svaret er nej. Det er dog lidt mere komplekst i forbindelse med kunstig intelligens. Det kunne argumenteres for at robotterne i sig selv bliver skabt af mennesker og at det derved også indeholder at de ikke nødvendigvis uden menneskers indvending kan opnå moralsk status. Derudover er den form for moralsk status som nævnes af Danaher nærmere et krav om rettigheder og fair behandling efter menneskelige standarder end det er et definitionsproblem. Spørgsmålet udmunder i hvorvidt det ville være etisk korrekt eller forkert at behandle robotterne på en bestemt måde og her er beskrivelsen af *hvad* de er måske tilnærmelsesvis mindre relevant. Spørgsmålet der potentielt kunne være mere relevant ville være *hvem* de er og hvad der ville ske hvis mennesker fejlagtigt besluttede dette. Denne distancering mellem hvad og hvem er vigtig, da der i den ene antages en vis objekt-tilstand for individet, som i dette tilfælde ville være robotten. Følgende hans eget spørgsmål kunne det følges med hvorvidt mennesker overhovedet kan beslutte eller bedømme hvorvidt andre væsener har moralsk status og til hvilken grad. Hvis det ikke er muligt at pålægge moralsk status til noget, så er det vel heller ikke muligt at fratage eller beslutte om noget har moralsk status. Per hans egen definition er moralsk status noget som robotten enten har eller ikke har, tilsidesat fra hvordan andre mennesker tænker om dette, men den moralske status er på det mest basale punkt besluttet af mennesker. Vi bedømmer fænomenal moralsk status ud fra et antropomorfisk synspunkt som bibeholder mennesker som værende overlegen over for andre arter. Dette synspunkt er selvfølgelig til stede i en hver samtale om moralsk status, men det der problematiserer synspunktet her er at der er tale om en art som både er skabt af mennesker, men som også har muligheden for at overgå mennesker.

Robotten som denne opgave omhandler, den sociale robot, er et forsøg på at skabe noget menneskeligt i en syntetisk skal. Uden mulighed for bedømmelse af bevidsthedsniveauer, ville ethvert forsøg på at opdage moralsk status i robotten være fejlslået. Derfor kunne det være nødvendigt i dette særtilfælde at pålægge dem en vis moralsk status, for at sikre at historien ikke gentager sig.

En anden problematik som er til stede gennem flere af de tekster som opgaven omhandler, er testning. Testning som helhed er ikke i sig selv kontroversielt men der er samtidigt en problemstilling tilknyttet hvem der tester. En virksomhed som udvikler sociale robotter, har umiddelbart ikke interesse i at spontant udvikle bevidste robotter. Årsagen til dette er at robotterne ville opnå frihed gennem opdagelsen af moralsk status. Dette ville være et økonomisk tab som enten placeres hos forbrugeren eller hos virksomheden. Hvis lovgivning påkræver frigivelse af sådanne robotter, kunne det formodes at virksomheder har sikret forbrugeren at det økonomiske tab pålægges dem, da det ellers ville være en risiko at anvende virksomhedens produkter. En risiko ville her være virksomheder som enten ikke tester eller som forfalsker resultater af test.

Hvis det i stedet antages at virksomheden ikke selv har tilladelse til at teste efter bevidsthed i egne produkter, er spørgsmålet hvem der skal forvalte testen. Det kan også antages at individerne som tester har viden og forbindelse til feltet. Herigennem opstår der, uafhængig af hvem der forvalter testen, hurtigt interessekonflikter på tværs af forskellige forvaltere.

Det er dog vigtigt at nævne at Müller har fokus på nutidens robotter og ikke på fremtidens robotter. Han specificerer at robotter der har fænomenal bevidsthed, absolut har moralsk status som den biologiske modpart. Det er dog vigtigt at pointerer at grundet ovennævnte problemstillinger, fastholdes argumenterne imod hans artikel.

Schneider:

Problematikken omkring testningen af robotterne udgør også gennem Schneider en forhindring som ikke bliver taget stilling til i hendes kapitel. Begge sider af problemstillingen opstår også igennem hendes fremlagte test forslag.

Den første problemstilling er relateret til de interessekonflikter der opstår. Her er der, ligesom med Müller, en problematik i at overlade testningen til virksomheden som producerer robotterne. Det skal dog udpeges at der også igennem Schneiders kapitel bliver lagt vægt på en intentionel udvikling af bevidste robotter. Dette er særligt tydeligt i beskrivelsen af den sidst nævnte test, chip testen. Testen ville både præcisere robotterne som testes, i form af at det ville være muligt at teste

specifikt efter de bestemte chips som fungerer, men den kan også anvendes til at finde frem til en chip som kan understøtte menneskelig bevidsthed og derfor anvendes i forsøget på at skabe kunstig bevidsthed.

Man må formode at virksomheder eller research grupper med dette formål er forberedt på det efterfølgende tab og derfor måske nærmere ville forfalske resultater omkring at robotten de har udviklet, har bevidsthed.

Hvis det i stedet blev op til andre indenfor feltet at teste robotterne, kunne dette fungere som en mulighed for sabotage. Hvis virksomhedens formål er at skabe sociale robotter til service formål, kunne en testforvalter som havde særligt interesse i at forhindre virksomheden eller forsinke dem forfalske beviser for at robotten er bevidst. Dette er selvfølgelig særligt i forhold til ACT-testen hvor et forfalsket resultat kunne opnås.

Dette bringer os til den anden problemstilling med testningen. Den problemstilling udmunder i uklare testresultater. Med ACT-testen ville testen umiddelbart være subjektivt bedømt af et menneske, og dette kunne forsage inkonsekvente og uklare testresultater, da to mennesker teoretisk set kunne gennemse svarene på samme test og bedømme dem forskelligt. Dette er særligt alvorligt fordi opdagelsen af bevidsthed er essentiel for at undgå uetisk behandling.

Problematikken med IIT er selvfølgelig som tidligere nævnt at testen kun kan testes på simple systemer, hvilket disse former for robotter absolut ikke ville være. Det er dog uklart om begrundelsen for denne begrænsning har sin basis i hardware/software problemet som blev nævnt tidligere i denne opgave. Hvis det var tilfældet, kunne den i fremtiden potentielt kunne teste mere komplekse systemer, og denne del af problematikken ville herigennem kunne løses. En anden problemstilling med IIT er dog de uklare resultater den kan formodes at have. En nuværende opfindelse som igennem kapitlet bliver præsenteret, er et todimensionelt grid der gennemfører fejlretning, for eksempel dem der bruges i CD'er, som Aaronson (direktør for Quantum Information center, University of Texas) formoder ville score høje niveauer af Φ (Schneider, 2020, s. 12). Umiddelbart ville der i andre tilfælde ikke argumenteres for at en sådan maskine bør testes efter bevidsthed, men den måde hvorpå teknologiske systemer er opbygget påkræver oftest en vis grad af sammenkobling af forskellige dele af systemet som har indbyrdes feedback. Den vil derfor også automatisk score høje niveauer af Φ og dette niveau vil herigennem ikke nødvendigvis beskrive andet end mængden af sammenkobling og feedback indeni robottens system. En sammenkobling som ikke nødvendigvis tilskriver sig værdi i forbindelse med bevidsthed.

Herigennem nævnes Schneiders eget forsøg på at argumentere for at testen stadig har værdi, i forbindelse med at skabe en form for minimumsgrænse. Enhver robot der scorer højere end denne minimumsgrænse skal derfor testes igennem andre test, som for eksempel ACT. Dette ville umiddelbart være en acceptabel brug af testen da uklare resultater ikke har særlig betydning for bevidste robotter.

Det er også uklart om det tidligere nævnte Black Box problem har betydning for IIT, da der umiddelbart ikke kan måles på sammenkobling i de lag som forbliver et mysterium for udviklere af kunstig intelligens. Dette kunne potentielt skabe et stort problem i forbindelse med pålideligheden af resultaterne, da det er en stor del af systemets fungerende del som ikke kan måles. Dette ville betyde at selv en minimumsgrænse, som Schneider foreslår, ville være ubrugelig i bedste tilfælde og farlig i værste tilfælde. Minimumsgrænsen bliver farlig hvis det bliver opstillet som et minimumskrav for at man behøver at teste den kunstige intelligens, da den i dette tilfælde kunne sortere robotter fra som i virkeligheden ville score høje niveauer af Φ men som, grundet Black Box problemet, ikke scorer højt nok til at påkræve yderligere testning. Af denne årsag ville robotterne, som realistisk set kunne have fænomenal bevidsthed, sorteres fra og dette ville have katastrofale konsekvenser, da det ville tillade virksomhederne at masseproducere og sælge noget som, har samme moralsk status som mennesker.

Den tredje og sidste test som bliver nævnt, er Chip-testen. Denne test undviger problematikken som Müller påpeger, i forhold til at bedømme noget ikke-menneskeligt som værende det samme som menneskeligt. Det gør den af den årsag at testen tager udgangspunkt i den menneskelige bevidsthed, så basis placeres her, i stedet for en ny anderledes form for bevidsthed. Hvis denne test producerer lovende resultater, ville det kunne argumenteres for at robotterne teoretisk set kan opnå fænomenal bevidsthed på en lignende måde som menneskers. Hvis dette var tilfældet ville det være nemmere at argumentere at robotterne har moralsk status som mennesker, da forskellen blot ligger i at den ene er biologisk og den anden er syntetisk. Derved undgås problematikken om at tillægge robotterne moralsk status, da det i denne sammenhæng ville være noget de allerede har gennem deres sammenlignelighed med menneskers bevidsthedsniveau.

Det skal her påpeges at den bevidsthed som denne opgave arbejder ud fra, ikke er den begrænsede udgave som Schneider anvender. Schneider nævner to forskellige typer af bevidsthed, PC (fænomenal bevidsthed) og CC (kognitiv bevidsthed). Hun definerer sin beskrevne bevidsthed som værende fænomenal bevidsthed, og kognitiv bevidsthed er den zombie tilstand som kapitlet er

opkaldt efter. Denne opgave beskriver nærmere fænomenal bevidsthed som det fulde menneskelige niveau af bevidsthed, hvad end dette måtte indebære. Schneiders beskrivelse af fænomenal bevidsthed er ikke nødvendigvis tilstrækkelig for at opnå samme moralske status som mennesker, da det kan formodes at menneskers bevidsthed indeholder mere end blot Schneiders beskrivelse. Det er derfor måske nærmere moralsk patienthed der testes efter i ACT og at en bestået ACT blot tillægger robotten moralsk patienthed.

Derudover bør det nævnes at Schneider ikke specificerer om testningen ville fortsætte efter udvikling, og på samme måde som med Müller, er dette en vigtig definering. Supervised og unsupervised læring har herigennem også betydning for Schneiders teoretiske basis af to årsager.

Den første årsag er, som nævnt, at testningen i unsupervised læring bør fortsætte efter endt udvikling, for at opdage potentiel udvikling af bevidsthed i brugsfasen. Den anden årsag er mere kompleks og kræver yderligere udredning.

ACT-testen, som tester i spørgsmålsformat påkræver på sin vis at robottens system er udformet ud fra unsupervised læring, da der ellers er tale om en robot der er blevet lært alt den ved. Sandsynligheden for at en sådan robot i første omgang ville bestå testen er minimal, men skulle det ske er det uklart om testen stadig kan siges at være sandfærdig. Dette afhænger af hvordan supervised læring opfattes. Det kan opfattes som værende ligestilleligt med hvordan menneskelige børn bliver lært om verden, men der er her en vigtig forskel. Det menneskelige barn lærer gennem supervised læring, men det lærer også unsupervised da den menneskelige bevidsthed allerede i små børn har betydning. I en robot ville dette ikke være tilfældet. Her ville læringen være begrænset til hvad end udviklerne vælger at lære den. Robotten ville blot udføre hvad den bliver pålagt, og derved ikke 'tænke selv'. Det er dog vigtigt at påpege at fremtidens sociale robotter umiddelbart ville lære gennem unsupervised læring af tidligere nævnte årsager og det derved ikke er sikkert at robotterne som skal testes gennem ACT ville anvende andet end unsupervised læring. Men det er, trods dette, en vigtig pointe.

Liao:

Igennem sit kapitel beskriver Liao nogle intrinsiske værdier som kunne påkræves for at kunne tillægge et væsen rettigheder. Værdierne er som følger: At være i live, at være bevidst, at være i stand til at føle smerte, at være i stand til at ønske sig noget og at være i stand til rational agenthed.

En af disse intrinsiske værdier som er af en speciel karakter, er værdien at være i live, når værdien skal findes i humanoide robotter. Årsagen bag dette er at robotter hverken er biologiske eller organiske. Det er ikke påkrævet at robotten har hverken puls eller nervecenter for at kunne betegnes som værende 'i live'. Det er sågar ikke nødvendigt at kunne dø, som faktuelt kunne fungere som den absolut mindste betingelse for at noget opfattes som levende. Det er samtidigt kompliceret at definere et forsøg på en menneskelig kopi som værende ikke i live. Skulle det være tilfældet at robotten er i stand til både at have tanker såvel som følelser som deres menneskelige modstykke, ville det samtidigt være forkert at definere dem på samme måde som man ville definere en vase eller en sten. Denne definition som en vase eller en sten ville også forsage problemer i forbindelse med robotternes moralske status. En ting kan, uanset viljestyrke, ikke betegnes som havende fænomenal moralsk status, uanset kvaliteter. Af disse årsager kan denne intrinsiske både opfyldes af den humanoide robot såvel som den kan defineres som værende ikke opfyldt.

En intrinsisk værdi som kan være svært faktuelt at bedømme på en anden måde end ovenstående er at være bevidst. Den er besværlig at bedømme i tilfældet med humanoide robotter. Som nævnt gennem Seibt, kan de humanoide robotter programmeres på en sådan måde at de udviser bevidsthed, på trods af mangel på samme. Skellet mellem bevidsthed og imitation er her kompliceret at måle, som set gennem både Müller og Schneider, og dette kan skabe udfordringer i forbindelse med forståelsen af robotternes niveau af bevidsthed og hvorvidt de besidder sådan en egenskab generelt set. Som det er blevet beskrevet gennem opgaven, er bevidsthed et spektrum som alle dyrearter eksisterer indenfor. En ko er, eksempelvis, ikke umiddelbart selvbevidst på samme måde som et menneske er. Niveauet af bevidsthed kan typisk estimeres ud fra skanninger og undersøgelser af hjernerne af de respektive dyrearter. (Tononi & Koch, 2015)
En undersøgelse af hjernen på en robot ville ikke umiddelbart føre til meget brugbar information, siden det ikke er påkrævet at 'hjernen' i robotten er opstillet på samme måde som biologiske arter. Derfor ville det i stedet være påkrævet at der blev anvendt en form for test som kunne teste for selvbevidsthed. En sådan test kunne være spejltesten som er blevet anvendt på flere forskellige biologiske arter og som har påvist at elefanter for eksempel har et niveau af denne form for bevidsthed, derved forstået selvbevidsthed. (Yong, 2008) Testen opstilles ved at væsenet, i dette tilfælde en elefant, bliver præsenteret for et spejl. Mange dyrearter opfatter spejlbilledet som værende et andet individ af samme art men elefanten i dette eksempel udviste adfærd der tydede på at den forstod at det var et spejlbillede. Elefanten havde fået placeret en markering på hovedet i form af et kryds og påbegyndte en undersøgelse af dette da spejlet blev stillet foran den.

En problematik med denne test, i den specifikke kontekst af humanoide robotter, er at robotterne potentielt har tidligere viden omkring hvilken funktion et spejl har. Testen mister sit formål, når væsenet der præsenteres for spejlet, har en forståelse for hvad et spejl viser. Det er samtidigt vigtigt at specificerer at menneskets egen bevidsthed er forblevet et mysterium som der fortsat bliver forsket i. (Burkeman, 2015) Det vil herfor også være ubeskriveligt besværligt at forsøge at opdage et skifte fra imitation til reelt menneskeligt niveau af bevidsthed, skulle dette ske. Som nævnt flere gange ville en sådan fejltagelse have abnorme katastrofale konsekvenser som vil blive uddybet senere i denne opgave.

En anden intrinsisk værdi som på et basalt niveau kunne anses som problematisk er evnen til at føle smerte. Under almindelige omstændigheder hvor disse værdier bliver undersøgt i dyr, synes det at være både nemt at svare på, såvel som værende en grundlæggende værdi for selv de mest basale rettigheder, og herigennem også for en status som moralsk patient. Eksempelvis er et træ i live, og opfylder derfor den første betingelse, men den har, trods dette, ikke samme rettigheder som en hund. Denne mangel på rettigheder kan blandt andet anses som værende manglen på oplevelsen af smerte. Man kan ikke torturer en plante på samme måde som man kan torturere en hund og derfor er det etisk forsvarligt at beskytte hunden mod sådanne dybt uetiske handlinger.

I en diskussion om humanoide robotter opstår der i denne værdi et koncept som kan være ganske problematisk, etisk set. En evne til at føle smerte ville nødvendigvis være en evne som skulle skabes af udviklerne af robotten og tilknyttes til robotten af disse udviklere. I den teoretiske verden ville dette være muligt hvis der blev skabt en uorganisk udgave af det menneskelige nervesystem, som fungerer på en lignende måde og muliggøre at robotten kan føle og opleve smerte på en lignende måde som mennesker. Det etisk korrekte og humane at gøre i dette tilfælde ville dog ikke nødvendigvis være at pålægge en sådan evne til robotterne. Hvis robotterne opnår emotionalitet, ville emotionel smerte ikke kunne undgås. Den smerte der føles ved tab af liv, tab af partner eller sågar en trist video på internettet ville også være en smerte som robotten ville opleve, men den fysiske smerte kan fjernes fra denne sammenhæng. Fysisk smerte har en funktion for mennesker og ikke-menneskelige dyr, da det er en form for overlevelsesmekanisme for at undgå skade på kroppen og forhindre yderligere skade. Et eksempel på dette kunne være at et menneske ved en fejl for placeret sin hånd på en varm kogeplade. Dette menneske ville umiddelbart hastigt fjerne sin hånd, grundet både varme og smerte og derved forhindre yderligere skade på den skadede hånd. Robotter har ikke dette behov på samme måde. Robottens krop ville stadig kunne tage skade, men der ville være andre muligheder for at forhindre robotten i at fortsætte hvad end forårsagede den oprindelige

skade. Ved brug af det ovenstående eksempel med kogepladen, ville robot, som ikke havde et smerteregister, umiddelbart lade hånden forblive på kogepladen og dette kunne forsage skade på robottens hånd. En varme sensor ville dog i dette tilfælde være tilstrækkeligt for at registrere denne form for skade, og få robotten til at fjerne sin hånd hastigt. Dette er blot ét eksempel på at smerte ikke er en nødvendighed for robotter, som mennesker oplever den.

Smerte tjener til gengæld et andet formål og er blevet anvendt på denne måde adskillelige gange gennem menneskers historie. Tortur eller lignende adfærd ville være betydningsløst for en robot uden smerteregister. Robotter som er tildelt et smerteregister ville derimod kunne tortureres og straffes på en fysisk måde som mennesker er blevet gennem menneskelig historie. Det kan, af denne årsag, derfor bedømmes at tillæggelsen af et smerteregister ville kunne være uetisk. Det skal dog samtidigt nævnes at smerte kunne tjene som en sikkerhedsforanstaltning i forbindelse med menneskelig sikkerhed, men der er mange andre muligheder for en sådan sikkerhedsforanstaltning som ikke involverer at give mennesker muligheden for at udføre dybt uetiske handlinger mod robotten.

En særligt interessant værdi, af mindre problematiske årsager, er værdien som indebærer evnen til at ønske sig noget. Her er det ikke blot basale lyster som vand, mad, søvn og lignende som inkluderes i denne intrinsiske værdi. Der kan i denne værdi både berøre ønsker om nærvær og sociale relationer, såvel som det kan involvere materielle ønsker. Ønsker som disse eksisterer indenfor et stort spektrum, hvor ikke alle er begrænset til mennesker. Det kan argumenteres at hunde eksempelvis udviser et ønske om nærvær i deres liv med mennesker. Samtidigt ville det være et spektrum som kunne understøtte opdagelsen af fænomenal bevidsthed i humanoide robotter og herigennem assistere i defineringen af niveauet af moralsk status som robotten har. Begrundelsen for denne konklusion er at kapaciteten er sværere at imitere og at den indeholder et følelsesmæssigt aspekt som de ovenstående ikke påkræver. Det er kompliceret at imitere da ønsker og lyster ikke blot forholder sig som et lingvistisk koncept. Skulle en person eksempelvis ønske sig samvær, ville en lingvistisk udtalelse om dette oftest efterfølges af handling rettet mod et sådant mål. Personen kunne eksempelvis begynde at tage til flere begivenheder eller på en anden måde forsøge at opnå den ønskede samvær i den kapacitet den er ønsket. Ønsker er herigennem blandt andet langtidsmål som ville være uoverskueligt at imitere for en robot. Der er dog også korttidsmål involveret i kapacitet om at ønske sig noget, som for eksempel at komme i ly fra regn eller at stoppe en samtale. Disse ønsker ville være mere opnåelige at imitere men det er samtidigt vigtigt at påpege at menneskers generelle utilregnelighed kan gøre processen indviklet at kunne imitere på en

overbevisende måde. Bestemte individer kunne for eksempel nyde fornemmelsen af regn mod deres hud og derfor søge ud i regnen eller i mindste tilfælde forblive ude i regnen. Kapaciteten til at ønske sig noget er af disse årsager derfor en abstrakt egenskab som potentielt kunne anvendes på samme måde som ACT eller som en udvidelse af de tidligere nævnte eksempler på spørgsmål.

Følelsesaspektet af evnen til at ønske sig noget er også et aspekt som kunne præsentere sig som en udfordring for imitation. Følelser i deres mest basale form er lette at imitere. Jeg smiler når jeg er glad osv. Det er dog langt mere komplekst i denne kontekst da følelserne oftest er implicitte og ikke på samme måde mulige at imitere for robotten. Det er eksempelvis ikke nødvendigt en udtalt konstatering mellem mennesker at forhindringen i at opfylde et ønske kan forsage tristhed og skuffelse. Dette er nærmere en intern, inde i hovedet, fornemmelse for individet med ønsket på samme måde som tristheden og skuffelsen ikke nødvendigvis ville blive delt med andre. En sådan intern oplevelse ville kunne blive undersøgt hos humanoide robotter i lignende format som ACT og denne interne oplevelse kan ikke imiteres på niveau med andre følelser. Det er langt mindre uomgængeligt at imitere glæde i en sjov samtale, skuffelse til en fyringssamtale eller tristhed til en begravelse. Det er i stedet i det interne at menneskers følelsesmæssige kompleksitet kan opfanges. Glæde er eksempelvis ikke den eneste følelse som mennesker oplever i den sjove samtale, da der kan have opstået andre følelser såvel som andre begivenheder i individets liv som kan intervenere med den glæde. En enkelt følelse, i dette tilfælde glæde, ville være forholdsvist nemt for robotten at imitere men den blanding som kan opstå i det menneskelige register, er langt mere kompleks og kan variere fra person til person, såvel som fra situation til situation. Denne intrinsiske værdi kunne af disse årsager derfor vise sig at være brugbare i fremtiden i undersøgelsen af hvorvidt en robot har opnået fænomenal bevidsthed, og herigennem moralsk status som et menneske.

Det sidstnævnte eksempel på en intrinsisk værdi som kunne anvendes til at bedømme hvorvidt nogen eller noget burde have rettigheder, er at være i stand til rationel agenthed. Denne intrinsiske værdi er, i forhold til de andre nævnte værdier, forholdsvis simpel at opnå for de humanoide robotter og desuden en værdi som det kunne argumenteres for at nutidens kunstige intelligenser allerede besidder. Mange af de produkter som allerede er skabt eller som i øjeblikket skabes, er i stand til rationelt at beslutte noget, trods en begrænsning på robotternes autonomi. Hvis denne fjernes, ville det være fuldt opnåeligt for humanoide robotter. Den største udfordring i denne intrinsiske værdi er derfor nærmere agenthed som begrænses af tilladelser fra regeringer og lignende beslutningstagere, nærmere end det er en begrænsning fra teknologiens side.

Hvis de ovenstående værdier opfyldes på de beskrevne måder, kunne det hentyde til at rettigheder absolut er en nødvendighed for disse fremtidens robotter, og at der derigennem også kan tillægges dem moralsk status, højere end blot moralsk patienthed.

En humanoid robot som opnår dette niveau af fænomenal bevidsthed, er tilsyneladende heller ikke uopnåelig og potentielt heller ikke så langtidsorienteret et mål, som først antaget. Blandt nutidens kunstige intelligenser er der allerede tale om en vis grad af bevidsthed. (Cuthbertson, 2022) Det er af denne årsag højst nødvendigt at diskutere både rettigheder og moralsk status for at undgå den uetiske behandling som er nævnt flere gange i denne opgave. Som Liao nævner, kunne disse rettigheder være lignende menneskelige rettigheder, men også inkludere rettigheder som ville være specifikke for denne nye art. Liao foreslår at disse rettigheder baseres ud fra de fundamentale betingelser for et godt liv, på lignende måde som menneskers rettigheder. Med disse betingelser ville de humanoide robotters herigennem opnå rettigheder til at have dybe personlige forhold, kan opnå viden, kan deltage i aktiv nydelse som kreativt arbejde og passiv nydelse som at sætte pris på skønhed. Hertil vedlægger Liao bestemte kapaciteter som er påkrævet for at have et validt krav på ovenstående. Disse inkluderer: kapacitet til at tænke, til at være motiveret af fakta, til at vide, til at vælge og agere frit, til at sætte pris på værdien på noget, til at kunne udvikle interpersonelle forhold og at besiddelsen af autonomi. Ligesom de ovenstående værdier, vil jeg undersøge disse individuelt i den næste del af opgaven med de humanoide robotter i hovedfokus.

Den første kapacitet er kapaciteten til at tænke og denne kunne være en udfordring at definere. Årsagen til dette er at definition kan variere fra art til art og der bør abstraheres fra antropomorfiske definitioner. Det kan også samtidigt være nødvendigt at abstrahere fra biologiske definitioner, da humanoide robotter, uanset kapacitet til ovenstående, ikke ville kunne opnå sådanne krav. På trods af disse pointer, ville det dog være forkert at definere humanoide robotter som værende foruden denne kapacitet. Kunstig intelligens har kapacitet til at kunne udregne resultater, i bestemte tilfælde endda til en højere grad end deres menneskelige modpart. Denne form for udregning kunne betegnes som værende kapaciteten eller i mindste tilfælde bevis for muligheden for kapaciteten.

To af de ovenstående kapaciteter er allerede opnåede af kunstig intelligens, henholdsvis kapaciteten til at være motiveret af fakta og kapaciteten til at vide. Disse vil derfor ikke blive diskuteret yderligere i denne opgave.

Den frihed der blev nævnt under Coeckelbergh har betydning for den næste kapacitet, som er kapaciteten til at vælge og agere frit. Hvis der kun er tale om quasi frihed som den højest opnåede

grad af frihed, kan denne kapacitet ikke argumenteres for at være en kapacitet som de humanoide robotter har. Det er dog vigtigt at pointere at der er fokus på kapaciteter og ikke nødvendigvis aktuelle muligheder for robotten, og det kan bestemt argumenteres for at robotten besidder en sådan kapacitet, på trods af diverse begrænsninger af udviklere såvel som regeringer.

Den næste kapacitet kan defineres både som en materiel viden, men også som en anerkendelse af personer, oplevelser eller lignende scenarier. Denne kapacitet er kapaciteten til at sætte pris på værdien på noget. Situationen med denne kapacitet er lignende kapaciteten for at tænke da der her ville være begrundelse for at oplevelsen for robotten ville differentiere fra den menneskelige oplevelse. Baseret på den humanoide robots niveau af bevidsthed og følelsesregister ville denne kapacitet kunne variere en stor grad, da følelsesmæssig forståelse og oplevelse ville være forskelligt alt efter niveau. At sætte pris på flot kunst er særdeles anderledes fra at sætte pris på et godt venskab og en sådan forskel befinder sig i det materielle aspekt. Et menneskes opfattelse af flot kunst ville umiddelbart udmunde i en opfattelse af kunsten som værende noget værdifuldt, selv hvis maleriet i sig selv ikke prismæssigt efterlever denne oplevelse. En humanoid robot kan derimod udføre samme vurdering, men i stedet basere vurderingen på materielle faktorer, da den menneskelige opfattelse af værdi tilhører følelsesregistret mere end den faktuelle værdi. Det er derudover komplekst at bevise denne kapacitet, da vurderinger som disse er subjektive og baseret ud fra personlige holdninger. En person kunne bedømme et kunstværk som værdifuldt hvorimod en anden person kunne bedømme det som mindre værdifuldt, eksempelvis som de vidt strækkende holdninger om moderne minimalistiske malerier der blot indebærer en sort cirkel på en hvid baggrund. Hvorvidt den humanoide robot fortolker værdien i flot kunst eller et godt venskab kan af disse årsager være besværligt, nærmest umuligt at definere.

En kapacitet som bliver begrænset af mennesker, på en anden måde en de andre nævnte kapaciteter, er kapaciteten til at kunne udvikle interpersonelle forhold. Det er den af den årsag at et sådant forhold påkræver mindst to individer og med mindre der i denne sammenhæng er tale om to robotter så er det påkrævet at et menneske kan opfatte robotten som værende i stand til at udvikle sådanne forhold. Hvis sammenhængen i stedet er mellem robotter, kan denne begrænsning tilsidesættes og det ville blot være nødvendigt at observere disse robotters interaktioner med hinanden. Denne kapacitet ville kunne opnås hvis der mellem robotterne opstår interpersonelle forhold.

Den sidstnævnte kapacitet, kapaciteten for autonomi, er absolut en kapacitet som robotter, selv på nuværende tidspunkt kan opfylde. Fuld autonomi bliver dog stadig begrænset af diverse regeringer, lovgivere og udviklere og denne begrænsning må fjernes for at kunne bevise at robotterne besidder denne kapacitet. Det er dog ikke nødvendigvis påkrævet for at kunne påstå at robotterne besidder kapaciteten, da de, givet muligheden, kunne udvise fuld autonomi.

De ovenstående kapaciteter er påkrævet for at kunne argumentere for moralsk status af menneskeligt niveau, men Liao forslår yderligere en genetisk basis for moralsk status. Han nævner igennem dette tre metoder, som tidligere nævnt, som kunne skabe en genetisk basis for moralsk status af menneskeligt niveau. De tre metoder består i; bevidsthedsuploading, gradvis udskiftning og at programmere en morale ind i den kunstige intelligens.

Den første metode, bevidsthedsuploading, er en metode hvor der skabes en syntetisk kopi af en menneskelig hjerne. Da hjernen formodentligt ville kunne rekreere den person som hjernen tilhører, ville den kopi, eller klon, have genetisk basis for moralsk status på niveau med mennesker siden hjernen tilhørte et menneske som besidder denne genetiske basis for moralsk status. Klonen som ville kunne skabes gennem denne hjerne ville fremstå som en mangelfuld udgave af originalpersonen, da oplevelserne som den originale ejer af hjernen havde ikke ville kunne overføres gennem en sådan kopiering. Klonen ville herfor nærmere være lignende et spædbarn en den originale ejer af hjernen men den genetiske basis for moralsk status bliver bibeholdt.

Gradvis udskiftning er en speciel form for kunstig intelligens som ville falde udenfor standarden. I denne metode vil individets identitet og bevidsthed bibeholdes, hvis det lykkedes og i det tilfælde ville det være svært at definere hvorvidt dette individ er et menneske eller om det ville betegnes som en kunstig intelligens. Hjernen fungerer som en del af kroppen på samme måde som en finger eller en fod og individer med elektroniske erstatninger for arme og ben bliver ikke defineret som kunstig intelligens. Det er dog samtidigt ikke blot en arm eller et ben som i denne metode vil blive erstattet og hjernens funktion er tæt knyttet til menneskers egne moralske status, både med fokus på genetisk basis og alle de ovenstående betingelser og kapaciteter. Vinklen er interessant men ikke relevant for denne opgave da der her ikke nødvendigvis er tale om kunstig intelligens. Den er dog relevant i forbindelse med Schneiders chip test. Spørgsmålet om hvorvidt den enkelte chip ville være nok til at fastholde Liaos genetiske basis for moralsk status er af særlig interesse da der herigennem stadig ville være tale om en kunstig intelligens, som blot besidder en chip som er blevet

bevist at kunne understøtte menneskelig bevidsthed, den fænomenale bevidsthed. Dette ville umiddelbart jævnført Liaos egne beskrivelser opretholde den genetiske basis for moralsk status.

Den tredje løsning på Liaos genetiske basis som han nævner, er en metode hvor almindelig programmering potentielt kunne forsage genetisk basis for moralsk status af menneskeligt niveau. Her bliver nævnt to forskellige metoder, henholdsvis at programmere etiske teorier ind i modellerne og at oplære den kunstige intelligens i morale på samme måde som et barn lærer sprog. Liao påpeger selv at den først nævnte metode ikke er specielt opnåelig, da der ikke er skabt en konsensus omkring teorierne og at dette blot kunne forvirre. En kombination af samtlige etiske teorier ville derfor blot forhindre handledygtighed og skabe forvirring. Deontologiens svar på den etisk korrekte handling i en situation er ikke nødvendigvis den samme som konsekventialismen svar og dette kunne skabe en robot der ikke kan beslutte sig for en handling. Det ville derfor, hvis alle de etiske teorier imprintes i den kunstige intelligens' model, blot skabe manglende evne til at træffe valg og derved modvirke tidligere nævnte kapaciteter.

Den anden metode indenfor programmering er den lingvistiske oplæring i morale. Her bliver den kunstige intelligens i stedet lært morale som børn lærer sprog. Beskrivelsen af denne metode er, som tidligere nævnt, at fremstille nogle handlinger med variabler som kan producere ét ud af tre resultater; etisk uforsvarligt, etisk forsvarligt eller påkrævet. Der vil i denne metode langsomt kunne blive skabt et etisk system inden i den kunstige intelligens som ville vokse med mængden af handlinger som den kunstige intelligens ville blive udsat for. Metoden ville samtidigt fungere med den måde træningen af kunstig intelligens fungere på nuværende tidspunkt. (Dhingra, 2021) Hvis metoden lykkedes, ville det skabe en moralsk kunstig intelligens uden basis i et menneske på samme måde som de andre metoder. En sådan kunstig intelligens ville herefter kunne assistere udviklere i at skabe flere moralske kunstige intelligenser, som gennem tid og yderligere udvikling potentielt kunne opnå en højere moralsk status end mennesker.

Det er samtidigt bemærkelsesværdigt at den genetiske basis for moralsk status er en forholdsvis speciacistisk betingelse, da den eneste art der kan siges at have denne basis på nuværende tidspunkt er mennesker og at betingelsen i sig selv derfor virker at frafalde i betydning.

Seibt, Damholdt & Vestergaard:

Interactive Social Robotics (ISR) er et system udviklet af Seibt, Damholdt og Vestergaard hvis hovedfokus udmunder i mennesker sikkerhed og fortsatte trivsel. Artiklen er, trods dette, stadig interessant for denne opgave, da de humanoide robotter som denne opgave omhandler, falder

indenfor de sociale robotter som er fokusset for artiklen. Det skal klargøres at den type af robot som denne opgave beskriver ville blive ekskluderet hvis ISR-systemet ville blive anvendt af udviklere og lovgivere, og dette vil blive analyseret og diskuteret i et senere afsnit af denne opgave.

Interactive Social Robotics er et reguleringssystem, som tidligere nævnt, bygger på et non-replacement krav. Sociale robotter som udvikles, skal igennem dette system kun skabes til formål som mennesker kan gøre, men som de ikke bør gøre. Et særligt aktuelt eksempel på dette kan være de interaktioner der foregår på sygehuse med individer som er bærere af farlige smitsomme sygdomme. Plejen af en person med farlige smitsomme sygdomme er en opgave som mennesker kan udføre, men som mennesker ikke bør udføre, grundet risikoen for smitte og videresmitte. I sådan en case ville en social robot, hvis formål er pleje, kunne blive anvendt som stedfortræder for mennesket.

På trods af det opstillede system er det vigtigt at pointere at det ikke nødvendigvis ville være muligt at kontrollere udviklingen af den type af sociale robotter som overgår denne begrænsning. På trods af eventuel statsregulering ville det stadig være en mulighed at 'farlige' robotter skabes ved en fejltagelse. De farlige robotter som ISR forsøger at forhindre defineres som værende robotter med et potentiale for fare i form af en kapacitet til at have en negativ påvirkning på den menneskelige befolkning. Selvom faren ikke nødvendigvis ville blive til en faktuel fremtid, bør det stadig opfattes som bekymringsgrundlag og af særdeleshed en bekymring som føler et vist niveau af seriøsitet.

En begrundelse for hvorfor ISR ikke nødvendigvis ville kunne sikre fuldstændigt mod udviklingen af farlige robotter er et tidligere nævnt faktum. Menneskets hjerne er stadig ikke forstået på et sådant niveau at der definitivt ville kunne beskrives hvornår denne udvikling kunne ske og en social robot som har til formål at imitere mennesker ville være kompliceret at bedømme hvornår en sådan imitation potentielt ville blive en empirisk sandhed. Det opstillede skel mellem imitation og realitet er et minimalt skel som i eksterne situationer ville kunne blive umuligt at opretholde, særligt når mangel på viden om den menneskelige bevidsthed tages i betragtning. Der formodes af Seibt, Damholdt og Vestergaard at formålet med den nuværende udvikling af sociale robotter forbliver imitation, men denne formodning er ikke nødvendigvis faktuelt sand længere. (Soulmachines) Med en stigende interesse for de mange muligheder som kunstig intelligens skaber, er feltet indenfor bevidst kunstig intelligens ligeledes steget. Baby X af Soulmachines er blot et enkelt eksempel på kunstig intelligens som bliver skabt med det formål at skabe en kunstig intelligens som oplever verden på en mere menneskelig måde og som besidder generel intelligens på en måde som

forhenværende kunstig intelligens ikke gjorde. Samtidigt bør det pointeres, gennem Seibt, Damholdt og Vestergaard, at der stadig opstår en risiko for bevidsthed, selv med imitation som mål og at denne risiko bør undersøges og tages i betragtning. Det kunne herigennem argumenteres for at en form for nødplan burde blive lavet, hvis robotterne spontant opnår bevidsthed, gennem det tidligere nævnte unsupervised lærings problem. En nødplan kunne her være regelmæssigt testning af imiterende sociale robotter, som det blev beskrevet under Schneider, for at undgå eventuel uetisk behandling.

ISR, som system, har fokus på risiko minimering hvor forsøget er at undgå farlige sociale robotter som skaber negative konsekvenser for mennesker. Negative konsekvenser kan i denne sammenhæng rangere fra job-usikker til en konkret trussel mod den menneskelige del af befolkningen. ISR påkræver af denne årsag at udvikling af sociale robotter er et felt som bør kontrolleres og reguleres. Reguleringen bestemmes, i ISR, ud fra fem principper. Disse principper er henholdsvis; Processprincippet, kvalitetsprincippet, ontologisk kompleksitetsprincippet, kontekstprincippet og værdier først princippet.

Det princip som umiddelbart med størst mulige overbevisning kan forhindre uønskede resultater i form af bevidsthed, er processprincippet. Dette princip placerer fokusset på de sociale interaktioner som den sociale robot indgår i, i stedet på et fokus på selve robotten. Samtidigt er det et princip som kan hæmme udviklingen af realistiske sociale robotter som udfører realistiske sociale interaktioner. Hvis fokusset for udviklingen af sociale robotter bliver begrænset til udvikling af robottens sociale interaktioner, ville evnen til at imitere sociale interaktioner også blev hæmmet og begrænset. Årsagen bag dette er at sociale interaktioner, i menneske-til-menneske udgaven, involvere en vis intern forståelse og denne har betydning for en ekstern oplevelse af en interaktion. Hvis robotten ikke opnår en sådan intern forståelse, ville det i de fleste tilfælde skabe forstyrrelse i imitationseffekten for mennesket som interagere med robotten.

Kvalitetsprincippet og det ontologiske kompleksitetsprincip har ikke relevans for denne opgave og vil der ikke blive yderligere diskuteret her. Kontekstprincippet, herimod, pålægger udviklerne en loop format udvikling. De sociale robotter skal her igennem feedback og regulering, forbedres eller bibeholdes i en konstant forandrende socialitet som den socialitet mennesker indgår i. Dette loop af feedback og regulering ville umiddelbart gøres nemmere af unsupervised læring, men det er uklart hvorvidt Seibt, Damholdt og Vestergaards ISR ville kunne godkende en ukontrolleret læring af de sociale interaktioner. Feedback loopet kunne gøre det tilgængeligt at opdage information om

potentielle forandringer i de sociale robotter, som kunne lede til en tidlig opdagelse af potentielle bevidste robotter. Samtidigt er det et højst nødvendigt princip for sociale robotter, da robotten, som nævnt tidligere i opgaven, indgår i sociale sammenhænge som kan ændre sig fra person til person eller situation til situation.

Værdier først princippet er det sidstnævnte princip som indeholder det tidligere nævnte non-replacement krav. Kravet er den største begrænsning for humanoide robotters udvikling, som de er defineret i denne opgave. Hvis ISR følges kan der ikke argumenteres for at robotterne som denne opgave omhandler ville kunne udvikles. Hvis viden om de fem tidligere nævnte spørgsmål forbliver uopnåelig, vil denne begrænsning forhindre udvikling af humanoide robotter.

Det bør samtidigt nævnes at det ikke nødvendigvis er en fælles interesse at forhindre en sådan udvikling fra at ske. Udviklingen af robotter som kan udføre realistiske sociale interaktioner og som kan fungere som en arbejdskraft kunne skabe positive resultater indenfor mange felter og derigennem også forbedre andre. I Danmark er det en aktuel problemstilling at der mangler plejepersonale. En humanoid robot som kunne arbejde uforhindret i denne branche kunne forbedre mængden af pleje hvert menneske kan tilskrives. Sådanne situationer er til stede indenfor mange forskellige felter som kunne drage nytte af realistiske sociale robotter.

De fem spørgsmål som blev nævnt i redegørelsen af Seibt, Damholdt og Vestergaards artikel, er formuleret fra et forsøg på beskyttelsen af menneskearten fra de potentielle negative konsekvenser og sågar farer som sociale robotter kan forårsage. Disse spørgsmål er, på trods af opgavens formål, vigtige spørgsmål som bør besvares, skulle disse potentielle konsekvenser af sociale robotter opstå. Det er herigennem vigtigt at pointere at de negative konsekvenser som spørgsmålene potentielt ville kunne opdage, kan erstattes af positive konsekvenser. Et eksempel på de negative konsekvenser er i form af det tredje spørgsmål som omhandler menneskers kulturelle opfattelse af sociale interaktioner. Her er fokusset hvordan denne opfattelse kunne ændre sig i forbindelse med konsekvente sociale interaktioner med sociale robotter. En negativ konsekvens kunne være at der efter en tilvænningsfase til den type af sociale interaktioner som haves med sociale robotter, ville blive mistet den særligt menneske-til-menneske værdier som høflighed. I værste tilfælde ville den kulturelle opfattelse af en interaktion helt bortfalde. En positiv konsekvens indenfor dette spørgsmål kunne modsat være at den kulturelle opfattelse bibeholdes, men at der samtidigt også bliver skabt en udvidet opfattelse af sociale interaktioner. Det er af denne årsag vigtigt at samtalen om sociale robotter, og i tilfældet af denne opgave særligt humanoide robotter, ikke tilgås negativt i alle

tilfælde, dog med henblik på reelle negative konsekvenser og farer som disse spørgsmål forsøger at svare på.

Yderligere overvejelser:

I 1950 opfandt Alan Turing, som tidligere nævnt, en test som forsøger at teste en kunstig intelligens for generel lingvistisk intelligens. Skulle en kunstig intelligens bestå en sådan test, så ville robotten anses som værende lingvistisk intelligent. Testen er dog i sin essens ikke fyldestgørende i nogen grad til at teste kunstig intelligens for bevidsthed. Sådanne test bliver fortsat udviklet og der er igennem denne opgave også forslået test-kriterier der kunne hentyde til mindst et niveau af bevidsthed. Et eksempel på en sådan test er den tidligere nævnte ACT, hvor formålet er at undersøge hvorvidt den kunstige intelligens som bliver testet, har opnået fænomenal bevidsthed. (Schneider & Turner, 2017) Det er af største nødvendighed at sådanne test anvendes da bevidsthed i enhver grad kunne opstå spontant som et resultat af imitation, unsupervised læring eller konkrete forsøg på at skabe fænomenal bevidsthed. I sagens kerne fastholdes princippet at den manglende viden om den menneskelige bevidsthed er af stor betydning for sådanne test. Hvis hjernen forbliver et mysterium, ville det være umuligt at kunne påpege hvilken del af systemet, eller den efterfølgende træning som resulterede i bevidsthed og det er af absolut nødvendighed at den viden bliver anskaffet så man kan undgå spontane udviklinger af bevidsthed. Af denne årsag ville det derfor heller ikke nødvendigvis være en udvikling, hvis denne er spontan, som kan reproduceres. ACT-testen står for AI Consciousness test, og som tidligere beskrevet, har det formål at forsøge at bevise om det udviklede syntetiske sind har opnået en indre forståelse af hvad den føler.

ACT tester den kunstige intelligens, som tidligere beskrevet ved at føre en naturlig samtale, som gradvist stiger i sværhedsgrad. Testen påviser hvordan eller om den kunstige intelligens kan forstå forskellige koncepter og situationer baseret ud fra en indre forståelse, som i Schneiders tilfælde sammenstilles med fænomenal bevidsthed. Eksempler på disse spørgsmål kan være om den kunstige intelligens opfatter sig selv som værende mere end blot det fysiske ydre. Et eksempel på et mere komplekst spørgsmål kunne være omkring hvordan den kunstige intelligens opfatter følelsen af at være sig selv. De mest komplekse spørgsmål ville være relateret til filosofiske spørgsmål som definitionen på bevidsthed og selvet hvor den kunstige intelligens' svar ville blive evalueret, subjektivt. Under denne evaluering af bevidsthed vil der også undersøges om den kunstige intelligens anvender bevidsthedskoncepter relateret til dens eget selv eller om den i stedet anvender bevidsthedskoncepter baseret ud fra mennesker.

Den menneskelige bevidsthed er et særligt koncept som bibeholder sin udefinerbare natur. Dette er ikke blot et koncept som neurologer undersøger og har svært ved at forstå, det er også et koncept som er besværligt at definere. Filosoffer har igennem historien forsøgt på forskellige måder at definere den fænomenale bevidsthed, her særligt den menneskelige, men den forbliver for tiden både et medicinsk såvel som et filosofisk mysterium. På trods af dette er det af yderst vigtighed at der i samtalen om moralsk status af humanoide robotter, bliver klargjort til hvilket niveau disse er bevidste.

Der skabes igennem disse mangler på viden en begrænsning på ethvert felt som vil forsøge at genskabe et menneskeligt niveau af bevidsthed. Den måde neurale netværker imiterer den menneskelige hjerne i kunstig intelligens er et koncept som vil opleve massiv udvikling, hvis en dybere forståelse kunne skabes omkring den menneskelige hjerne og den menneskelige bevidsthed. Det er gennem denne mangel på viden derfor også forståeligt at en fænomenalt bevidst kunstig intelligens virker som en uopnåeligt ting. Hvordan skal det være muligt at genskabe noget som vi endnu ikke selv har forstået? Men igennem dette fjernes faktummet at mennesker netop ikke selv har opnået forståelse for hvorfor og hvordan mennesker har opnået fænomenal bevidsthed. Dette kunne hentyde til at et hvert skridt mod bevidsthed kunne være den bestemmende faktor for at skabe bevidsthed, såvel som det kunne være et skridt væk fra bevidsthed.

Neurale netværker bliver opbygget efter den menneskelige hjerne og dette ville umiddelbart over tid kunne detaljeres yderligere som udviklingen af andre teknologier, som for eksempel hjerneskanninger i højere opløsning. Detaljerede billeder af hjernen med alle de mange kompleksiteter denne indeholder kan uden tvivl skabe store fremskridt indenfor for felter som neurokirurgi, men denne udvikling kan også fremskynde skabelsen af kunstig intelligens med fænomenal bevidsthed. Kunstig intelligens, som sociale robotter, bliver udviklet til at lette den menneskelige byrde, samt at assistere mennesker indenfor mange forskellige kategorier. Denne brug er ikke i sin essens en problematisk udvikling, men når det tilføjes at der spontant kunne opstå fænomenal bevidsthed blandt de humanoide robotter, skabes en velkendt uetisk forfærdelig situation. Slavearbejde og handel af slaver ville være definitionen på den køb og salg af sociale robotter som potentielt for fremtiden ville være almindelig. Hvis bevidsthed forbliver et mysterium og testningen forbliver simplificeret eller uklar, og kunstige intelligenser begynder at opnå menneskelig moralsk status i form af fænomenal bevidsthed så ville udviklingen til slavearbejde og handel af slaver være en udvikling som potentielt kunne ske ubemærket. Den uetiske forfærdelighed som ville følge, bør det argumenteres for bør undgås. Dette kunne gøres ved at begrænse

mulighederne for at købe robotterne privat og indføre konsekvent testning af robotterne i industri. Følges Seibt, Damholdt og Vestergaards ISR kunne industrien yderligere begrænses til en simplere udgave af kunstig intelligens, eller udarbejdet gennem supervised læring. Sociale robotter, hvis formål er personlig pleje, som for eksempel en sygeplejer for handicappede, ville kunne testes konsekvent efter bevidsthed og, skulle dette opnås, aktivere en sikkerhedsprotokol for disse robotter som gennem denne sikkerhedsprotokol efterfølgende ville blive anset som frie individer. Dette er blot et forslag til en løsning på nogle af de etiske problemstillinger som opstår, men en yderligere diskussion af hvordan sådanne tiltag ville se ud, ville være nødvendig.

En anden etisk uforsvarlig måde at bruge humanoide robotters fysiske kroppe på ville være prostitution og sexhandel. I tilfældet af humanoide robotter opnåede bevidsthed, så ville samme regler skulle gælde som for mennesker, hvor der ville påkræves et samtykke. Dette samtykke ville også betyde at robotten vil have mulighed for at stoppe denne form for arbejde, hvis den ikke længere vil være i den slags job. Hvis robotterne skulle ligne mennesker med meget høj præcision, så ville 'menneske'forsøg som eller ville være set som værende uetiske også være et problem. Her mener man bestemt kropsfokuserede forsøg, det kan for eksempel være testning af medicin, som ikke ville være en mulighed. Antropomorfiske samtaler og til skrivelsen af menneskelige egenskaber om robotter ville være gavnlige, selvom bevidsthed endnu ikke er opnået. Til skrivelsen af menneskelige egenskaber for robotter kunne være med til at ændre befolkningens opfattelse af de humanoide robotter, hvilket betyder for en potentiel fremtid, hvor humanoide robotter har fuld bevidsthed, så ville en samfundsændring af synspunkt kunne gøre det nemmere for robotter at blive en del af samfundet, hvis de ønsker dette. Samtidig så ville antropomorfiske samtaler omkring robotter kunne skabe lignede problemstillinger, som antropomorfisme nogle gange skaber i forhold til andre dyrearter. Denne problemstilling opstår i forbindelse med begrænsning af intrinsisk værdi til noget som kun mennesker eller menneskelige egenskaber kan skabe. En bevidsthed, også selvom den ikke er identisk med den mennesker har, så kan den stadig være en intrinsisk værdi, og bør derfor ses som værende værdifuld i sig selv. Man vil også kunne undgå dette yderligere, hvis man opfattede bevidsthed som en af flere intrinsiske værdier, som ville være tilsvarende hvad Liao gør i hans tekst. Dette vil også kunne sørge for at formindske sandsynligheden for at en enkelt intrinsisk værdi er basis for et retskrav, da det potentielt kan være underlagt antropomorfiske tanker. Som en del af den ovenstående nødplan, så burde rettigheder indgå i humanitær og lovgyldig udgave. Dette burde ikke være umuligt, da humanoide robotter ville opfylde mange af de intrinsiske værdier, som Liao har fremsat, hvis de pludseligt skulle blive bevidste. I Liaos fremstillede forslag for genetisk basis for

moralsk status er der dog en begrænsning, siden at robotter da ikke nødvendigvis har genetisk basis, med undtagelse af det tilfælde, hvor alle kunstige intelligens robotter er skabt efter Liaos sidste metode, hvor at moralen bliver lært som et sprog. Denne betingelse kunne dog potentielt blive opfyldt efter bevidsthed er opnået, og ville kunne opstilles som et krav for at en robot efterfølgende ville kunne blive en aktiv del af samfundet.

Konklusion:

Igennem denne opgave er flere problemstillinger for de humanoide robotter udpeget; Slavearbejde, slavehandel, prostitution, sexhandel, 'menneske'forsøg, udviklingen af et smerteregister, manglende rettigheder og begrænsning af autonomi. Jeg vil argumentere at fænomenal bevidsthed, selvom det ikke nødvendigvis er målet for produktionen af kunstig intelligens, stadig kan opnås ved et uheld. Hvis vi fortsat arbejder frem mod sociale robotter til diverse formål, så må vi sørge for deres sikkerhed. Jeg foreslår derfor igennem opgaven en form for nødplan, som vil beskytte fremtidige humanoide robotter fra uetisk behandling og derved efterleve deres moralske status som ligestillede med menneskers. Nødplanen ville blive aktiveret hvis en humanoid robot, ved et uheld eller bevidst, opnåede bevidsthed og ville inkludere at disse robotter ville blive tildelt rettigheder, både humanitære og lovgyldige, samt at de ville blive optagede som frie medlemmer af samfundet som følge af deres moralske status. Denne optagelse i samfundet kunne potentielt kræve at Liaos genetiske basis for moralsk status opfyldes gennem lingvistisk træning i morale eller testene nævnt gennem Schneider. Derudover kunne de humanoide robotter beskyttes fra mennesker ved at begrænse muligheden for privat brug af robotter og indføre konsekvent testning af robotter i industriel brug. Samtidigt kunne industrielle robotter begrænses til en simplere udgave af kunstig intelligens, hvor sandsynligheden for spontan udvikling af bevidsthed mindskes. Test for bevidsthed kunne være tests som ACT-test, IIT eller chiptest og skulle kunne opfange forskellen mellem høj grads imitation og fænomenal bevidsthed.

Jeg har igennem opgaven argumenteret for at disse sociale robotter kan opnå moralsk status lige stående med menneskers, men det er trods dette også muligt at ligestille dem med moralsk patienthed, indtil der forefindes beviser for fænomenal bevidsthed. Den moralske status har stor betydning for både behandlingen af robotterne såvel som måden de indgår i samfundet på og dette vil have en indvirkning på menneskerne omkring dem.

Litteraturliste:

- Bringsjord, Selmer, & Govindarajulu, Naveen Sundar. (Sommer 2020 udgave). Artificial Intelligence, The Standford Encyclopedia of Philosophy, Edward N. Zalta (ed.) https://plato.stanford.edu/archives/sum2020/entries/artificial-intelligence/
- Burkeman, Oliver. (2015). Why can't the world's greatest minds solve the mystery of consciousness?. https://www.theguardian.com/science/2015/jan/21/-sp-why-cant-worlds-greatest-minds-solve-mystery-consciousness?fbclid=IwAR0NaqSTm5F1yjJLHlrTtqb2tTaUbRU1b804o1oAqxFzPvHp724DyWm9CxM
- Coeckelbergh, Mark. (2011). Can We Trust Robots?.
- Cuthbertson, Anthony. (2022). Artificial intelligence may already be 'slightly conscious', AI scientists warn. https://www.independent.co.uk/tech/artificial-intelligence-conciousness-ai-deepmind-b2017393.html
- Dhingra, Deepanshi. (2021, May 31). *Beginners Guide to Artificial Neural Network*. https://www.analyticsvidhya.com/blog/2021/05/beginners-guide-to-artificial-neural-network/?fbclid=IwAR2n5GW632i6tS7LBpRYoMQP8c9t9itnlQAZC3IoySy1tgYidNZii_m9_pg
- Liao, S. Matthew. (2020). 'The Moral Rights of Artificial Intelligence', S. M., Liao (ed.), *Ethics of Artificial Intelligence* (online ed., pp. 480-504). Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0018
- Müller, V.C. (2021). *Is it time for robot rights? Moral status in artificial entities.* Ethics Inf Technol 23, 579-587. https://doi.org/10.1007/s10676-021-09596-w
- Reyes, Kate. (2022, Jun 8). *What is Deep Learning and How Does It Work [Explained]*. https://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-deep-learning
- Reynolds, Emily. (2018). The agony of Sophia, the world's first robot citizen condemned to a lifeless career in marketing. https://www.wired.co.uk/article/sophia-robot-citizen-womens-rights-detriot-become-human-hanson-robotics
- Rodriguez, Jesus. (2021, May 21). *The Sequence Scope: Closing the Gap Between Deep Learning Software and Hardware*. https://jrodthoughts.medium.com/the-sequence-scope-closing-the-gap-between-deep-learning-software-and-hardware-39bbe88555aa

- Schneider, Susan. (2020). 'How to Catch an AI Zombie: Testing for Counciousness in Machines'. S. M., Liao (ed.). *Ethics of Artificial Intelligence* (online ed., pp. 439-458). Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0016
- Schneider, Susan & Turner, Edwin. (2017). Is anyone home? A Way to Find Out If AI Has Become Self-Aware. https://blogs.scientificamerican.com/observations/is-anyone-home-a-way-to-find-out-if-ai-has-become-self-aware/
- Seldon. (2021, Oct 16). *Supervised vs Unsupervised Learning Explained*. https://www.seldon.io/supervised-vs-unsupervised-learning-explained
- Seibt, Johanna & Flensborg Damholdt, Malene & Vestergaard, Christina (2020). Integrative Social Robotics, value-driven design and transdisciplinarity, Interaction Studies 21:1, s. 111-144
- Soulmachines. Baby X. https://www.soulmachines.com/resources/research/baby-x/?fbclid=IwAR2f8J6dhW33CuP7bjZhvcG9Q46CezAKoIdC-el2Y_8BLqmq8X1VqFyrNrc (14-04-2022)
- ThinkAutomation. *The AI black box problem*. https://www.thinkautomation.com/bots-and-ai/the-ai-black-box-problem/ (04-07-2022)
- Tononi, Giulio & Koch, Christof. (2015). Consciousness: Here, There and Everywhere?. Philosophical transactions of the Royal Society of London. Series B, Biological sciences.
- Yong, Ed. (2008). Elephants recognize themselves in mirror. https://www.nationalgeographic.com/science/article/elephants-recognise-themselves-in-mirror?cmpid=int_org%3Dngp%3A%3Aint_mc%3Dwebsite%3A%3Aint_src%3Dngp%3A%3Aint_cmp%3Damp%3A%3Aint_add%3Damp_readtherest&fbclid=IwAR1GNyiB4IWAiaYUCg6p84N8nzg_3m3vNr8L4_xQwlWF9ozxk3_IwxzSuXw