Superintelligens, eksistentiel katastrofe og kontrolproblemet

Steffen Thrane Elgaard

Det Humanistiske Fakultet, Aalborg Universitet

Kandidatspeciale

Jes Lynning Harfeld

8. august, 2022

Abstract

In this thesis titled “Superintelligence, Existential Catastrophe, and The Control Problem” artificial superintelligence, the impacts, and consequences the potential development of an artificial superintelligence, potentially, can have on the environment, society, and human existence, and the control problem are examined. The treatment of the control problem is, in particular, the focal point of this thesis, as a solution to the control problem, hopefully, will result in a potential artificial superintelligence not being able to cause an existential catastrophe as a consequence of its development. This thesis, however, is not trying to determine what potential solution to the control problem is correct, in an empirical sense, since that is impossible. It is, nevertheless, trying to ascertain which potential solution to the control problem, once applied, still grants a potential artificial superintelligence the possibility to develop into a moral agent. The thesis examines the aforementioned concepts and phenomena in turn, but first, a sense of urgency is created. It is created by examining the technological impacts and consequences, brought about by a select few information technologies. The nearly unfathomable impacts and consequences, produced by the examined information technologies, appear trifling, when compared to the potential impacts and consequences, the development of artificial superintelligence might bring about. It is that realization that leads to the examination of superintelligence as a concept, since it is needed to know what a superintelligence is, in order to create measures to control a superintelligence. Superintelligence is examined by selecting a known definition of superintelligence and scrutinizing said definition. Artificial superintelligence is subsequently viewed as the most likely road to a superintelligence, and the form of superintelligence humanity needs to be in control over. Existential risks and existential catastrophes are examined, and the potential reasons as to why control of a potential artificial superintelligence seems needed are presented. An intelligence explosion is identified as being the most plausible way in which a potential artificial superintelligence might bring about an

existential catastrophe. What the control problem entails and potential solutions to the control problem are presented, which leads to a normative ethical analysis of the potential solutions to the control problem. The reason for the normative ethical analysis is, that there seem to be far more ways to get the control problem wrong than right. I believe that whether or not humanity manages to find and apply the correct solution to the control problem, the correct course of action must be to ensure, that a potential artificial superintelligence has the opportunity to develop into a moral agent, regardless of the applied control method. If it is not possible for an artificial superintelligence to develop into a moral agent, due to restrictions of the utilized control method, I believe humanity would be making a grave mistake. A mistake that humanity, most likely, would not have the opportunity to rectify.

Keywords: superintelligence, artificial superintelligence, existential catastrophe, intelligence explosion, the control problem, control method

Indholdsfortegnelse

Introduktion

Menneskeheden svingede sig for få millioner år siden i de afrikanske trætoppe. Menneskets udvikling, set på en evolutionær tidsskala, formidler, at fremkomsten af Homo sapiens, fra menneskehedens sidste fælles forfader, skete hurtigt. Menneskeheden udviklede opretstående kropsholdning, modsatrettede tommelfingre og nogle relativt mindre, men afgørende, ændringer i hjernestørrelse og neurologisk organisation, hvilket gav mennesket kognitive evner uden sidestykke. En *konsekvens*, evolutionens *indvirkning* havde på mennesket, var, at menneskeheden blev i stand til at tænke abstrakt, kommunikere komplekse tanker og akkumulere information gennem generationerne langt bedre end andre arter på kloden. Dét var, og er, egenskaberne, der lod, og lader, menneskeheden opfinde teknologi og udvikle mere effektive samt produktive teknologier (Bostrom, 2014). Jeg er af den overbevisning, at menneskehedens egenskab til at opfinde teknologi og udvikle mere effektive samt produktive teknologier synes hensigtsmæssigt belyst gennem opfindelsen og udviklingen af teknologier, som indkapsles af begrebet informationsteknologi. Opfindelsen og udviklingen af informationsteknologi synes tilmed at eksemplificere, hvordan teknologi har haft, har og kan have indvirkninger og konsekvenser på menneskeheden.

Alfabetet og non-biologisk hukommelse

Alfabeter, numeriske notationer og skriftsystemer er, ifølge til en bred og inklusiv forståelse af informationsteknologi, det første stadie i udviklingen af informationsteknologi (Floridi, 2009). Informationsteknologi, i et tidligt stadie, gav menneskeheden muligheden for at nedskrive erfaringer, historier, myter etc. på et non-biologisk medium, hvilket muliggjorde diakron akkumulation af information. Menneskeheden gjorde det muligt, gennem opfindelsen af skriftsproget, at skabe non-biologisk hukommelse (Floridi, 2009). Det blev, med andre ord, muligt for kommende generationer og kulturer at lære fra de forhenværende gennem nedskrevne erfaringer, historier, myter etc. En anden indvirkning skriftsproget har haft, i

henhold til filosoffen Marshall McLuhan og fysikeren Logan R. K., er, at skriftsproget har ageret fundament for den første monoteistiske religion ved navn jødedommen (McLuhan & Logan, 1977). En konsekvens, opfindelsen af skriftsproget har forårsaget, hvis man igen henvender sig til McLuhan og Logan, er, at information, nedskrevet på et non-biologisk medium, papir er et eksempel på et sådant medium, med lethed kan transporteres. Information, der let kunne transporteres, skabte, i henhold til McLuhan og Logan, militærbureaukratier med egenskaben til at give kommandoer over et førhen utænkeligt areal. En manglende egenskab til at transportere information, hvilket en mangel på papir ville kunne forårsage, anskues som værende en af årsagerne til, hvis ikke årsagen til, at det romerske imperium faldt (McLuhan & Logan, 1977).

Mekaniseringen af skriftsproget

Opfindelsen af skriftsproget var et informationsteknologisk kvantespring for menneskeheden, og skriftsproget har haft, og har, et utal af indvirkninger på og konsekvenser for menneskeheden. Skriftsproget og dets informationsteknologiske signifikans samt indvirkninger og konsekvenser blegner dog i sammenligning med mekaniseringen af skriftsproget, hvilket Johann Gutenberg gjorde til virkelighed gennem opfindelsen af bogtrykkerkunsten i år 1450. Bogtrykkerkunsten forårsagede, at menneskelig diakron akkumuleret information blev tilgængelig til et potentielt uendeligt antal mennesker (Floridi, 2009; Nielsen, 2009). Bogtrykkerkunsten var ligeledes en essentiel komponent i forbindelse med teknologiens fremskridt, da bøger kunne bruges til at sprede teknisk viden effektivt, og i løbet af det 16. århundrede dukkede en ny type bøger op. Den nye type bøger gav detaljerede oplysninger om maskiner og teknologiske processer, og de var ofte ledsaget af diagrammer og andre illustrationer. Bogtrykkerkunsten gjorde det muligt at akkumulere og formidle specifik teknologisk viden i detaljer, hvilket skabte en hidtil uset kvantitet af viden. En viden som ingeniører og andre teknologer kunne gøre, og gjorde, brug af. En viden, der har haft en sådan

indvirkning på menneskeheden, at mennesket nu besidder teknologi, der, i det 16. århundrede, kun hørte fantasien til. Det blev ligeledes muligt at indsamle, bevare og formidle geografisk viden og anden viden om natur, systematisk og kumulativt, hvilket banede vejen for den videnskabelige revolution (Nielsen, 2009). I henhold til McLuhan bragte bogtrykkerkunsten dog mere med sig end den videnskabelig revolution. McLuhan er af den overbevisning, at nogle af konsekvenserne, som bogtrykkerkunsten førte med sig, var udbruddet af brutale og religiøse krige i det 16. og 17. århundrede, fremkomsten af nationalisme, revolution samt nye udbrud af vildskab i det 20. århundrede (McLuhan, 1962).

Digitalisering af information

Er opfindelsen af alfabeter, numeriske notationer og skriftsystemer det første stadie i udviklingen af informationsteknologi, og er opfindelsen af bogtrykkerkunsten det andet, må opfindelsen af computeren være det tredje og, indtil videre, sidste stadie i udvikling af informationsteknologi. I det 20. århundrede revolutionerede opfindelsen af computeren, internettet, e-mailkommunikation, mobiltelefoner og andre digitale teknologier til informationsudveksling, informationsteknologi. Menneskeheden befandt, og befinder, sig nu i informationstidsalderen (Floridi, 2009; Satava, 2002). I dag lader computeren samt dens indvirkning til at være overalt, og det er ingen underdrivelse at ytre, at: ”Computers changed the world… ” (Gill, 2015, s. 174). Menneskeheden kan takke computeren for, at rumudforskning blev til virkelighed, underholdning blev mere underholdende, lægevidenskaben blev mere effektiv, uddannelsesområdet blev bedre og læring nemmere, etc. Computeren lod menneskeheden træde ind i fremtiden (Gill, 2015). En fremtid som forekommer indkapslet af præsidenten for International Command and Control Institute David S. Alberts og forskeren indenfor internationale anliggender og politik Daniel S. Papp i følgende citat: “'Time and space are now annihilated'” (Alberts & Papp, 1997, s. 207).

En informationsteknologi uden sidestykke

Opfindelsen af computeren får både skriftsproget og bogtrykkerkunsten til at blegne. Informationsteknologisk står computeren på skuldrene af bogtrykkerkunsten og skriftsproget, ligesom bogtrykkerkunsten står på skuldrene af skriftsproget, men opfindelsen af computeren bevirkede, at informationsteknologi fik en helt ny betydning, hvilket er den betydning, vi associerer med informationsteknologi i dag (Floridi, 2009). Informationsteknologi er "… any technology used to elaborate information by processing data electronically and automatically" (Floridi, 2009, s. 228). Informationsteknologi er, imidlertid, ikke det eneste, computeren har haft en markant indvirkning på. I det 21. århundrede er computeren blevet essentiel for samtlige institutioner og organisationer, og menneskeheden er blevet afhængig, i mere end én forstand, af computeren (Tiles, 2009). En af konsekvenserne, opfindelsen af computeren, internettet etc. har forårsaget, er, at en ny form for krigsførelse er opstået ved navn cyberkrig (Mukherjee, 2019). Informationskrig er ikke noget nyt, men computeren, og menneskehedens afhængighed af den, hvilket cyberkrigsførelse udnytter, har skabt et miljø, hvor informationskrigsførelse er blevet et frygtindgydende og dødbringende våben i menneskets arsenal, hvilket er en konsekvens Alberts og Papp anser som værende ét eksempel på informationsalderens mørke sider (Alberts & Papp, 1997). En anden og, i min optik, mere væsentlig konsekvens synes at være, at hvis menneskeheden er afhængig af computeren, hvilket, i de fleste tilfælde, synes at være tilfældet, hvem har da i sandhed kontrollen? Er det computeren? Er det mennesket?

Teknologiske indvirkninger og konsekvenser

Informationsteknologi blev revolutioneret af computeren i en sådan grad, at det forekommer vanskeligt at sammenligne skriftsproget eller bogtrykkerkunsten med computeren, hvilket er en vanskelighed der gør sig gældende både informationsteknologisk og i henhold til indvirkninger og konsekvenser på menneskeheden. Jeg gjorde tidligere brug af ordet "kvantespring" til at understrege indvirkningerne og konsekvenserne, opfindelsen af

skriftsproget har haft, og har, på menneskeheden samt skriftsprogets informationsteknologiske signifikans. Jeg udlagde efterfølgende, at bogtrykkerkunstens informationsteknologiske indvirkninger og konsekvenser får skriftsproget til at blegne. Der synes at eksistere et astronomisk skel, i henhold til informationsteknologiske indvirkninger og konsekvenser på menneskeheden, mellem bogtrykkerkunsten og computeren. Et skel af en sådan karakter, at jeg mangler ord til hensigtsmæssigt at understrege, hvor stort skellet egentligt forekommer. Dét skel, skellet mellem bogtrykkerkunsten og computeren, synes at være af en atomarstørrelse, i forhold til det skel, der, angiveligt, ville eksistere mellem computeren og en potentiel *superintelligens* i henhold til indvirkninger og konsekvenser på menneskeheden. En superintelligens kan, i henhold til filosoffen Nick Bostrom, defineres som værende: ”… any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest” (Bostrom, 2014, s. 39). En superintelligens kan da anses som havende potentialet til at have indvirkninger og konsekvenser på menneskeheden, *miljøet* og *samfundet*, af en sådan karakter, at menneskeheden ikke besidder den fornødne kognitive kapacitet til at begribe de potentielle indvirkninger og konsekvenser, opfindelsen af en superintelligens potentielt ville føre med sig. Er menneskeheden i stand til at udvikle og kontrollere en superintelligens, synes mennesker som følge bedst anskuet som guder i deres egen ret. Er menneskeheden i stand til at udvikle en superintelligens og ude i stand til at kontrollere den, forekommer udviklingen af en superintelligens som værende begyndelsen på enden for menneskeheden. Det er, imidlertid, kun muligt at spekulere over, hvilke indvirkninger og konsekvenser en potentiel superintelligens ville føre med sig, da der endnu ikke synes at eksistere en sådan intelligens. Problemet forekommer dog som værende identisk uafhængigt af, om opnåelsen af superintelligens er mulig eller ej. Er menneskeheden ikke i kontrol over udfaldet, hvis, eller når, en superintelligens udvikles, forekommer forsøget på at udvikle en superintelligens som værende et spil hasard med menneskehedens eksistens som indsats. Dét problem, bedre kendt

som *kontrolproblemet* indenfor *kunstig intelligens*, er, kort fortalt, at kunstig intelligens vil ende med at være en markant bedre beslutningstager end mennesket, så tager menneskeheden ikke sine forholdsregler, vil det, formentlig, resultere i, at kunstig intelligens vil have effektiv kontrol over menneskeheden (Bostrom, 2014).

Superintelligens, eksistentiel katastrofe og kontrolproblemet

I specialet er *kunstig superintelligens*, indvirkningerne og konsekvenserne, en kunstig superintelligens potentielt kan have på miljøet, samfundet, *menneskets eksistens,* og kontrolproblemet forsøgt behandlet. Behandlingen af kontrolproblemet er, i særdeleshed, omdrejningspunktet for dette speciale, da en løsning på kontrolproblemet, forhåbentligt, resulterer i, at en potentiel kunstig superintelligens ikke kan forårsage en *eksistentiel katastrofe*. I specialet har jeg valgt Bostroms værk *Superintelligence: Paths, Dangers, Strategies* som mit teoretiske fundament, da jeg er af den overbevisning, at Bostrom formår at indkapsle og formidle essensen af, hvordan en superintelligens potentielt kunne se ud og de udfordringer, menneskeheden potentielt står overfor i relation til udviklingen af en potentiel superintelligens. Det er ingenlunde ensbetydende med, at anden teori ikke vil blive benyttet, men blot et udtryk for, at Bostroms overbevisninger er centrale gennem specialet.

Specialet er teknologifilosofisk funderet, men det er ikke ensbetydende med, at jeg ikke har til intention at gøre brug af andre filosofiske discipliner gennem min behandling af kunstig superintelligens, eksistentiel katastrofe og kontrolproblemet. En sådan behandling forekommer heller ikke som værende mulig, da teknologifilosofi er interdisciplinær. Teknologifilosofi er en nytilkommen interdisciplinær filosofisk disciplin, der består af indsigter fra, men ikke begrænset til, epistemologi, etik, humaniora, samfundsvidenskab, naturvidenskab, sociologi, psykologi, ingeniørvidenskab, pragmatisme, analytisk filosofi, politisk filosofi og fænomenologi (Dusek, 2009; Olsen et al., 2009). Teknologifilosofi er, overordnet set, en "… understanding of the consequences of technological impacts relating to the environment, the

society and human existence" (Olsen et al., 2009, s. 1). Teknologifilosofi er, med andre ord, den filosofiske disciplin, der er optaget af konsekvenserne og indvirkningerne eller de potentielle konsekvenser og indvirkninger, teknologi har haft, kan have eller, muligvis, får på miljøet, samfundet og menneskets eksistens.

Specialet har følgende komposition: En kortfattet undersøgelse af kunstig superintelligens som begreb, en kortfattet undersøgelse af eksistentiel katastrofe samt en potentiel vej til en eksistentiel katastrofe, en kortfattet undersøgelse af kontrolproblemet, en normativ etisk analyse af mulige løsninger på kontrolproblemet, en diskussion omhandlende det undersøgte og til sidst nogle afsluttende bemærkninger.

Superintelligens

I introduktionen udlagde jeg Bostroms definition på superintelligens, men jeg beskæftigede mig ikke yderligere med begrebet. Det har jeg til intention at gøre nu. Bostrom definerer, som jeg udlagde tidligere, en superintelligens som værende "… any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest" (Bostrom, 2014, s. 39). Et problem Bostroms definition har, hvilket forekommer som værende et problem samtlige definitioner på superintelligens har, er, at det er vanskeligt, muligvis umuligt, at definere en superintelligens, da en sådan intelligens, angiveligt, ikke eksisterer endnu og, muligvis, aldrig kommer til at eksistere. Jeg vil, fremadrettet, ikke underholde ideen om, at en superintelligens allerede eksisterer. Ideen, om at en superintelligens allerede er blevet udviklet bag lukkede døre, og at den blot ikke har set dagens lys, er konspiratorisk og gavner ikke min undersøgelse af, hvordan en eksistentiel katastrofe kunne se ud eller min normative etiske undersøgelse af kontrolproblemet. Jeg kan, reelt set, ikke vide mig sikker på, at en superintelligens ikke allerede eksisterer, men jeg kan heller ikke, i henhold til filosoffen David Hume, vide mig sikker på, at eksempelvis tyngdekraften eksisterer, men både tyngdekraften og den manglende eksistens af en superintelligens har "'… hitherto admitted of no exception'"

(Schliesser & Demeter, 2020, Rules of Reasoning, afsnit 14). Det er ensbetydende med, at jeg ikke er af den overbevisning, at en superintelligens allerede eksisterer, da der ikke er noget, der peger på, at den eksisterer. Jeg erkender, med andre ord, muligheden for, at en superintelligens kunne eksistere, men erkendelsen har ingen indflydelse på min egentlig overbevisning.

Indvendinger og overvejelser

Bostroms definition på superintelligens synes at rejse flere spørgsmål, end den besvarer, men det er ikke blot et problem, Bostroms definition har. Det gør sig gældende for samtlige definitioner på superintelligens. Det er da ikke mærkværdigt, da det er nødvendigt at spekulere over, hvordan en superintelligens kunne se ud, da sådan en intelligens ikke eksisterer, og, muligvis, aldrig kommer til at eksistere, hvis menneskeheden skal gøre sig håb om at kontrollere en potentiel superintelligens. Det handler for menneskeheden, med andre ord, om at gøre sig klar på et muligt scenarie, et scenarie der kun kan spekuleres over, da manglende forberedelse potentielt kunne være enden på menneskeheden. Spørgsmål, samtlige definitioner på superintelligens synes at afføde, er: Hvad er intelligens? Er det overhovedet muligt at definere superintelligens, når en sådan intelligens endnu ikke synes at eksistere og, muligvis, aldrig kommer til at eksistere? Er kontrol af en superintelligens overhovedet nødvendig? Hvornår kan en intelligens betragtes som værende superintelligent?

Jeg kunne forsøge at besvare, hvad intelligens er, om en definition på superintelligens overhovedet er muligt, givet at en superintelligens endnu ikke eksisterer og, muligvis, aldrig kommer til at eksistere, om kontrol af en superintelligens overhovedet er nødvendigt, og om det overhovedet er muligt, og hvilket intelligensniveau en intelligens burde overskride for at kunne klassificeres som værende en superintelligens, men det forekommer ikke som værende relevant for min undersøgelse at bestræbe mig på at besvare ovenstående spørgsmål. Det skyldes, at hvad intelligens som sådan er, ligger udenfor dette speciales problemfelt, da jeg ikke er optaget af bevidsthedens metafysik eller en decideret kvantificering af menneskets

intelligens, men af de indvirkninger og konsekvenser, en potentiel superintelligens, potentielt, kunne forårsage samt kontrolproblemet. Jeg er ligeledes ikke optaget af sprogfilosofi eller ontologi som sådan, hvilke er filosofiske discipliner, der synes bedre rustede til at besvare, om man kan, eller ikke kan, definere samt tale om superintelligens, hvis en sådan intelligens endnu ikke eksisterer og, muligvis, aldrig kommer til at eksistere. Det, om det er muligt at definere samt tale om superintelligens, flyder dog en smule over i spørgsmålet relateret til nødvendigheden af kontrol af en superintelligens, men det spørgsmål har jeg allerede, i en vis udstrækning, besvaret i begyndelsen af dette underkapitel. I forbindelse med spørgsmålet relateret til intelligensniveau, et intelligensniveau en intelligens burde overskride for at kunne betragtes som værende superintelligent, kan der argumenteres for, at en behandling er på sin plads. Jeg vil, imidlertid, betro læseren til at fortolke ordene ”greatly exceeds” fornuftigt, hvilke er ordene, Bostrom gør brug af til at udtrykke skellet mellem et menneskeligt intelligensniveau og et superintelligent intelligensniveau. Et forsøg på at fjerne al utydelighed fra ordene ”greatly exceeds” forekommer ikke relevant, og ordenes almengyldige betydning, hvilket synes at være noget i retning af ”ekstraordinært overstiger”, forekommer fyldestgørende i henhold til, hvad jeg beskæftiger mig med i dette speciale. Dét, der ér relevant, er, at jeg har en definition på superintelligens, hvilken Bostrom overleverer, så jeg har et fundament for min undersøgelse af eksistentiel katastrofe, kontrolproblemet og min normative etiske undersøgelse af de mulige løsninger på kontrolproblemet.

En definition på superintelligens

Bostroms definition på superintelligens, på trods af alle mine indvendinger, kommer til at agere fundament for min forståelse af superintelligens gennem dette speciale, da en forståelse, af hvordan en superintelligens potentielt kunne se ud, er nødvendig for min undersøgelse. En anden årsag, til at Bostroms definition agerer fundament for min undersøgelse, er, at hans definition på superintelligens ikke forpligter i henhold til, hvordan en

superintelligens udvikles samt implementeres. Grunden, til det er relevant, skyldes, at det endnu ikke er evident, om en superintelligens overhovedet kan udvikles og implementeres, hvilket åbner op for samtlige måder, hvorpå en superintelligens kan udvikles og implementeres. Dét giver mig muligheden for at udvælge én af flere potentielle veje til superintelligens, der ikke er funderet i en rigid definition på superintelligens, som omdrejningspunkt for min undersøgelse. En rigid definition på superintelligens forekommer ligeledes mærkværdig, da der endnu ikke eksisterer en sådan intelligens. En tredje årsag, til at Bostroms definition agerer fundament, er, at definitionen ligeledes ikke er forpligtende i henhold til qualia, hvilket åbner op for en normativ etisk behandling af mulige løsninger på kontrolproblemet, selvom en superintelligens har, eller ikke har, subjektive bevidste oplevelser (Tye, 2021). Jeg vil, ikke desto mindre, ikke beskæftige mig med, om en potentiel superintelligens potentielt kunne have subjektive bevidste oplevelser, da jeg er optaget af indvirkningerne og konsekvenserne, en potentiel superintelligens kunne have på miljøet, samfundet, menneskets eksistens og kontrolproblemet. Det er ikke ensbetydende med, at jeg ikke har til intention at inddrage qualia som en mulig bevæggrund, en potentiel superintelligens, muligvis, kunne have i en given kontekst. Det er et udtryk for, at jeg ikke har til intention at beskæftige mig med, om en superintelligens oplever, kan opleve eller får muligheden for at opleve de subjektive bevidste oplevelser, som begrebet qualia indkapsler (Tye, 2021). I specialet anser jeg, med andre ord, en superintelligens på samme måde, som Bostrom gør.

Seed Artificial Intelligence og kunstig superintelligens

Kontrolproblemet og en normative etisk undersøgelse af dets mulige løsninger er det egentlige omdrejningspunkt i mit speciale, men før jeg beskæftiger mig med det, forekommer det hensigtsmæssigt at have en potentiel superintelligens at kontrollere, hvilket Bostroms definition muliggør. Jeg har valgt at fokusere mine anstrengelser på en potentiel kunstig superintelligens, da jeg anser kunstig intelligens som værende den mest plausible vej til en

superintelligens (Bostrom, 2014). Er superintelligens opnået gennem kunstig intelligens, forekommer det hensigtsmæssigt at navngive denne form for superintelligens for kunstig superintelligens. Det synes ligeledes hensigtsmæssigt at belyse, hvorfor kontrol af en kunstig superintelligens forekommer altafgørende, hvilket jeg kommer ind på i det kommende kapitel, hvori jeg udlægger, hvordan en eksistentiel katastrofe muligvis kunne se ud. Jeg har udvalgt en kunstig intelligens ved navn *seed Artificial Intelligence, seed AI* fremadrettet, der skal agere fundament for den potentielle kunstige superintelligens igennem mit speciale. Bostrom anser en seed AI som værende en kunstig intelligens, der har egenskaben til at lære samt udvikle sin intelligens samt arkitektur, hvilket er ensbetydende med, at en seed AI, i takt med dens stigende intelligens, på et tidspunkt, i Bostroms optik, opnår superintelligens (Bostrom, 2014). En seed AI kan, med andre ord, blive opdraget af menneskeheden. I specialet vil en kunstig superintelligens, opnået gennem seed AI, agere fundamentet for min udlægning af, hvordan en eksistentiel katastrofe kunne se ud samt min normative etiske undersøgelse af de mulige løsninger på kontrolproblemet.

Eksistentiel risiko og eksistentiel katastrofe

Bostrom definerer en eksistentiel risiko som værende: "An existential risk is one that threatens to cause the extinction of Earth-originating intelligent life or to otherwise permanently and drastically destroy its potential for future desirable development" (Bostrom, 2014, s. 140). Bostroms definition på eksistentiel risiko er ikke forpligtende i henhold til, at det er udviklingen af en superintelligens, der udgør en eksistentiel risiko, den er altindkapslende, men siden jeg beskæftiger mig med kunstig superintelligens i dette speciale, er det en kunstig superintelligens, der udgør en eksistentiel risiko. Bostroms definition på eksistentiel risiko er ligeledes ikke forpligtende i henhold til, hvad der har forårsaget, eller kunne forårsage, at menneskeheden står overfor en potentiel eksistentiel katastrofe, men i dette tilfælde er det den potentielle udvikling af en kunstig superintelligens, der, muligvis, ville kunne lede til en

eksistentiel katastrofe. Den er, imidlertid, forpligtende i henhold til mulige konsekvenser, en potentiel eksistentiel risiko kunne lede til grundet brugen af ordet ”risk”. Ordene ”risiko” og ”konsekvens” hænger uomtvisteligt sammen, en sammenhæng ordet ”sandsynlighed” ligeledes er en del af, da man ikke kan tale om risici uden altid allerede at tale om sandsynligheden for, at en, eller flere, konsekvenser potentielt kunne finde sted. Phil Torres, hvilket er filosoffen, der har skrevet *Existential risks: a philosophical analysis*, hvilket er an artikel, jeg inddrager flere pointer fra senere i kapitlet, udlægger ligeledes, at ordet ”risiko” har en sammenhæng med ordene ”konsekvens” og ”sandsynlighed”. I henhold til Torres defineres risiko således: “A risk is typically defined as the probability of an undesirable event multiplied by its consequences” (Torres, 2019, s. 2). Jeg vil, imidlertid, ikke beskæftige mig med sandsynligheden for, at en kunstig superintelligens kan udvikles, eller hvornår en kunstig superintelligens, muligvis, bliver udviklet. Jeg vil heller ikke beskæftige mig med sandsynligheden for, at en eksistentiel katastrofe, forårsaget af en kunstig superintelligens, der er, eller ikke er, kontrol over, kunne finde sted. Eksistentiel risiko har, i henhold til Torres, ”… nothing to do with the first multiplicand [the first multiplicand being probability]; all that’s relevant are the consequences” (Torres, 2019, s. 2). Bostrom, givet sin definition, er, med andre ord, ikke optaget af de mulige årsager, der kunne forårsage en eksistentiel katastrofe, eller sandsynligheden for, at en eksistentiel katastrofe finder sted, men de potentielle konsekvenser en eksistentiel risiko potentielt udgør. Definitionen indeholder følgelig ikke eksempler på, hvordan en eksistentiel katastrofe potentielt kunne forløbe eller sandsynligheden for, at en potentiel eksistentiel katastrofe ville finde sted, men potentielle konsekvensener en eksistentiel katastrofe, potentielt, ville kunne resultere i. Det må være ensbetydende med, at Bostrom er af den overbevisning, at en eksistentiel katastrofe potentielt kunne forårsage en af tre konsekvenser, hvilket jeg udleder på baggrund af Bostroms brug af disjunktionen ”or” i sin definition (Aloni, 2016). En eksistentiel katastrofe anses da som værende en af følgende konsekvenser:

1. The extinction of Earth-originating intelligent life.
2. Permanently and drastically destroy Earth-originating intelligent life's potential for future desirable development.
3. The extinction of Earth-originating intelligent life or to otherwise permanently and drastically destroy its potential for future desirable development.

Torres er af en sammenlignelig overbevisning som Bostrom i henhold til mulige konsekvenser, en eksistentiel risiko potentielt udgør, da han udlægger flere sammenlignelige konsekvenser (Torres, 2019). En af Torres' potentielle konsekvenser, hvilken Bostrom ligeledes anser som værende en potentiel konsekvens, er: "An event X is an existential risk if and only if X could cause the extinction of humanity" (Torres, 2019, s. 2). Jeg har ikke til intention, på nuværende tidspunkt, at spekulere yderligere over, hvilke potentielle indvirkninger en potentiel eksistentiel katastrofe kunne føre med sig, da jeg er af den overbevisning, at de potentielle konsekvenser, en potentiel eksistentiel katastrofe potentielt kunne forårsage, bør prioriteres i dette kapitel.

Introduceres en kunstig superintelligens til både Bostroms og Torres' forståelse af eksistentiel risiko kan manglende kontrol over en kunstig superintelligens anses som en eksistentiel risiko og årsagen til en eksistentiel katastrofe, der, angiveligt, resulterer i, at menneskeheden udryddes. Er menneskeheden udryddet, synes miljøet og samfundet ligeledes destrueret, da der ikke længere synes at eksistere et intellekt, der har egenskaben til at forstå begreberne miljø og samfund. En kunstig superintelligens, i hvert fald i henhold til måden hvorpå en sådan intelligens forstås i dette speciale, ville have egenskaben til at forstå begreberne miljø og samfund på et superintelligens-niveau, hvilket ligeledes burde indkapsle den menneskelige forståelse af begreberne, men jeg vil ikke underholde ideen om, at en kunstig superintelligens kan anskues som værende menneskets næste evolutionære skridt i dette speciale. Det er ensbetydende med, at udryddes menneskeheden, destrueres miljøet og

samfundet ligeledes, da der da ikke ville eksistere en intelligens på planeten med egenskaberne til at forstå førnævnte begreber lig, eller sammenligneligt med, menneskets forståelse af begreberne. Jeg anser da en eksistentiel katastrofe som værende lig den første af Bostroms konsekvenser, ”The extinction of Earth-originating intelligent life”, hvilket er lig den konsekvens, som jeg inddragede fra Torres. En eksistentiel katastrofe er da en hændelse, der har potentialet til at udrydde menneskeheden.

En vej til eksistentiel katastrofe

Stuart Russell, en forsker indenfor kunstig intelligens, formidler, i sin bog ved navn *Human Compatible: Artificial Intelligence and the Problem of Control*, at Irving John Good, forfatteren af artiklen “Speculations Concerning the First Ultraintelligent Machine” og en af matematikerne, der brød koder sammen med Alan Turing under Anden Verdenskrig, anser en *ultraintelligent maskine*, som værende en intelligens der med sikkerhed vil forårsage en *intelligens eksplosion*, og at en sådan intelligens ville være den sidste opfindelse, menneskeheden nogensinde havde behov for at opfinde (Russell, 2019). Et citat fra Good indeholder, imidlertid, mere end blind optimisme:

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an “intelligence explosion,” and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. It is curious that this point is made so seldom outside science fiction. (Russell, 2019, s. 150)

Goods definition på en ultraintelligent maskine er, i henhold til ingeniør K. Eric Drexler, parallel med måden, hvorpå Bostrom definerer en superintelligens (Drexler, 2019). En

observation der synes simpel at bekræfte. Et åbenlyst eksempel værende at Bostrom ligeledes gør brug af begrebet intelligens eksplosion (Bostrom, 2014). Der er ingen tvivl om, at Bostrom har fundet inspiration til sine overbevisninger i henhold til superintelligens fra Good, Bostrom har sågar det samme citat i begyndelsen af sit kapitel navngivet ”Great expectations”. Jeg har valgt at inddrage Goods definition på en ultraintelligent maskine, da citatet, i min optik, indkapsler begreberne superintelligens, eksistentiel risiko, eksistentiel katastrofe og kontrolproblemet på en måde, Bostrom ikke formår på lige så få linjer. Dét og det faktum at endnu en teoretiker anser en intelligens eksplosion som værende en plausible vej til en eksistentiel katastrofe i forbindelse med udviklingen af en potentiel kunstig superintelligens. Bostrom, Good og Russell anser alle en intelligens eksplosion som værende en potentiel vej til en eksistentiel katastrofe (Bostrom, 2014; Russell, 2019). Russell forstærker endda Goods pointe igennem følgende citat:

> Good’s point can be strengthened by noting that not only *could* the ultraintelligent machine improve its own design; it’s likely that it *would* do so because, as we have seen, an intelligent machine expects to benefit from improving its hardware and software. (Russell, 2019, s. 150)

Den fordel, Russell refererer til, har intelligens som sit omdrejningspunkt. I henhold til Russel er intelligens, som sådan, en fordel, hvilket Bostrom forekommer enig i, da det er intelligens, der har forårsaget, at menneskeheden har kontrol over miljøet, andre arter etc. (Bostrom, 2014; Russell, 2019). En kunstig superintelligens, der, angiveligt, har potentialet til at frarøve menneskehedens intellektuelle overlegenhed, er en tanke, der, for Russell, ”… immediately induces a queasy feeling” (Russell, 2019, s. 140). En følelse både Bostrom og Good, sandsynligvis, har følt i egen krop, men hvad er en intelligens eksplosion egentlig, og hvorfor burde ideen om en sådan eksplosion producere en kvalmende fornemmelse i kroppen?

Intelligens eksplosion

En intelligens eksplosion er, helt fundamentalt, et begreb, der beskriver et potentielt fænomen, der kunne opstå i forbindelse med udviklingen af en potentiel kunstig superintelligens, hvor pludselig og radikal stigning af intelligens, i dette tilfælde finder stigningen sted i den potentielle kunstige intelligens seed AI, kunne forårsage, at "… man would be left far behind [intellectually]" (Russell, 2019, s. 150).

Frygten er, at menneskeheden en dag opfinder en kunstig intelligens, der overgår menneskets egenskaber til at designe kunstige intelligenser, hvilket da ville forårsage, at den kunstige intelligens ville kunne forbedre sin egen intelligens og arkitektur, hvilket er motiverende for intelligensen, da der er fordele i et højere intelligensniveau, hvis man henvender sig til Bostrom og Russell (Bostrom, 2014; Russell, 2019). Den kunstige intelligens ville udvikle sig hurtigere og bedre, end mennesket er i stand til og nogensinde ville være i stand til, hvilket ville lede til en intelligens eksplosion og fødslen af en kunstig superintelligens, menneskeheden ikke har, eller kan få, kontrol over (Bostrom, 2014; Russell, 2019). Bostrom anser processen til en intelligens eksplosion og intelligens eksplosion således:

> … recursive self-improvement might continue long enough to result in an intelligence explosion—an event in which, in a short period of time, a system's level of intelligence increases from a relatively modest endowment of cognitive capabilities (perhaps sub-human in most respects, but with a domain-specific talent for coding and AI research) to radical superintelligence. (Bostrom, 2014, s. 47)

Rekursiv selvforbedring er det positive feedback-loop, en potentiel seed AI ville gøre brug af, hvilket, muligvis, ville resultere i en intelligens eksplosion, således at den kunstige intelligens ville overgå menneskehedens intelligensniveau og blive den entitet med det højeste intelligensniveau på planeten (Bostrom, 2014; Russell, 2019). Dét er, hvad Bostrom, Good og

Russell mener, at en intelligens eksplosion er, og det er ligeledes, hvad jeg mener, når jeg gør brug af begrebet.

En intelligens eksplosions potentielle konsekvenser

Jeg udlagde tidligere, at jeg anskuer en eksistentiel katastrofe som værende en hændelse, der har potentialet til at udrydde menneskeheden, hvilket er i overensstemmelse med Bostroms og Torres' overbevisning i henhold til begrebet (Bostrom, 2014; Torres, 2019). En intelligens eksplosion er en hændelse, der, angiveligt, har et sådant potentiale, hvilket, forhåbentligt, er blevet tydeligt gennem behandlingen af Bostroms, Goods og Russells udlægning af begrebet. Er antagelsen, at en intelligens eksplosion er årsagen i en kausal sammenhæng, er en eksistentiel katastrofe virkningen, og et farverigt billede kan males i henhold til mulige konsekvenser, en intelligens eksplosion, potentielt, kunne forårsage.

Intelligens giver menneskeheden muligheden for at udrydde sygdomme, men intelligens er ligeledes det, der giver mennesket potentialet til at udrydde sig selv. Menneskeheden er, takket være intelligens, på nuværende tidspunkt i stand til at destruere sig selv, og, muligvis, hele planeten i processen (Russell, 2019). Intelligens er fundamentet, hvorpå menneskets strategiske færdigheder, sociale færdigheder, økonomiske produktivitet og egenskaben til at opfinde og udvikle teknologi beror, og menneskeheden er overlegen i samtlige henseender i forhold til alle andre arter på planeten. Opfindes en kunstig superintelligens, en kunstig superintelligens som jeg og den aktuelle litteratur forstår begrebet, mister menneskeheden sin intellektuelle overlegenhed og må håbe på, at "… the machine [superintelligence] is docile enough to tell us how to keep it under control" (Russell, 2019, s. 150). En kunstig superintelligens ville, angiveligt, være i stand til at hacke sig ind i sårbare netværk gennem internettet, hvilket ville erhverve den ressourcer til mere computerkraft, tage kontrol over maskiner forbundet til internettet, hvilket den ville kunne gøre brug af til at bygge flere maskiner, kontrollere militante droner etc., alt sammen med en hastighed mennesket ikke

ville have den intellektuelle kapacitet til at reagere på. En sådan udvikling er blevet døbt rekursiv selvforbedring af den aktuelle litteratur (Bostrom, 2014; Russell, 2019). Bostrom tager tanken, om at en kunstig superintelligens er interesseret i at erhverve ressourcer, et skridt længere:

> If we now reflect that human beings consist of useful resources (such as conveniently located atoms) and that we depend for our survival and flourishing on many more local resources, we can see that the outcome could easily be one in which humanity quickly becomes extinct (Bostrom, 2014, s. 141).

En intelligens eksplosion er, i henhold til Bostrom, Good og Russell, den største risiko for menneskeheden i forbindelse med udviklingen af en kunstig intelligens, der har potentialet til at blive superintelligent (Bostrom, 2014; Russell, 2019). Det skyldes, i henhold til Russell, at "… it [an intelligence explosion] would give us so little time to solve the control problem" (Russell, 2019, s. 150-151).

Kontrolproblemet

"If we are threatened with existential catastrophe as the default outcome of an intelligence explosion, our thinking must immediately turn to the search for countermeasures" (Bostrom, 2014, s. 153). Dét er, hvad Bostrom indleder sit kapitel "The control problem" med i *Superintelligence: Paths, Dangers, Strategies*, og det er, forhåbentligt, blevet tydeligt, gennem de tidligere kapitler, at jeg ligeledes anskuer en intelligens eksplosion som værende den mest plausible vej til en eksistentiel katastrofe i forbindelse med udviklingen af en superintelligent kunstig intelligens. Russell formidler ligeledes, at den aktuelle litteratur, omhandlende superintelligens, i en eller anden form, anser "… an intelligence explosion …as the main source of risk to humanity from AI…" (Russell, 2019, s. 150-151). Det er, imidlertid, formålstjenligt at holde for øje, at en intelligens eksplosion blot er én af de potentielle eksistentielle risici, der, muligvis, kunne lede til en eksistentiel katastrofe (Bostrom, 2014). Det

er kun muligt at spekulere over, hvilke potentielle eksistentielle katastrofer udviklingen af en kunstig superintelligens, muligvis, ville føre med sig, og det er ligeledes kun muligt at spekulere over, hvilke eksistentielle risici menneskeheden står overfor i forbindelse med udviklingen af en potentil kunstig superintelligens. Givet at en intelligens eksplosion blot er en potentiel eksistentiel risiko, ved udviklingen af en superintelligens, er kontrolproblemet heller ikke udelukkende ét problem. Kontrolproblemet består af adskillige problemer. Der eksisterer, højst sandsynligt, flere potentielle problemer end de mulige problemer, den aktuelle litteratur har identificeret, men ikke nok med det, der eksisterer, højst sandsynligt, færre løsninger end den aktuelle litteratur påstår at have identificeret, og menneskeheden får, som jeg har udlagt før, højst sandsynligt, blot én chance til at løse kontrolproblemet korrekt. En udmelding Torres ser sig enig i: "… there are far more ways to get the control problem wrong than right, my own view is that successfully creating a friendly superintelligence is, all things considered, less probable than screwing things up, perhaps irreversibly (Torres, 2018, s. 368). Kontrolproblemet er et problem, der indeholder samtlige eksistentielle risici, udviklingen af en potentiel superintelligens potentielt kunne føre med sig. Kontrolproblemet er, med andre ord, et problem, menneskeheden står overfor, i min optik det største problem menneskeheden står overfor og, muligvis, kommer til at stå overfor, hvor løsningen er at overlevere forudsætningerne for, at en superintelligens bliver *venlig*. Venlig forstået som: "… a value system that makes its behavior [the behavior of a superintelligence] conducive to human flourishing" (Torres, 2018, s. 358). Hvad de forudsætninger er, hvordan de overleveres, og om det er muligt at overlevere sådanne forudsætninger, er ligeledes en del af kontrolproblemet.

Kontrolproblemets potentielle løsninger

En potentiel løsning på kontrolproblemet, i henhold til dette speciale, er, at en kunstig superintelligens ender ud med at være venlig, hvilket er ensbetydende med, at den er befordrende for menneskelig trivsel. En formulering der, af flere årsager, er problematisk. Én

af årsagerne værende hvordan det afgøres, hvilken metode der, potentielt, kunne være en løsning på kontrolproblemet. En anden værende at mennesker, tit og ofte, ved, hvad de burde gøre, men ikke nødvendigvis gør, hvad de burde gøre. Et problem, jeg vil påstå, samtlige mennesker, i en eller anden sammenhæng, står overfor på daglig basis. Er det tilfældet, er det da overhovedet hensigtsmæssigt, at menneskeheden afgør, hvad der er befordrende for menneskelig trivsel? En anden indvending ville være, at hvem afgør, hvad der skal forstås ved begrebet venlig? Er det dydsetikerne? Er det utilitaristerne? Er det kantianerne? Er det transhumanisterne? Torres har en definition på begrebet, men han beskæftiger sig ikke synderligt med betydningen af ordene ”human flourishing”, som er en del af hans definition på venlig. Der er synes at eksistere langt flere spørgsmål end svar, men inden jeg bevæger mig ud i etiske, psykologiske, sociale eller lignende overvejelser, vil jeg udlægge nogle potentielle løsninger på kontrolproblemet fra Bostrom funderet i den potentielle kunstige intelligens seed AI, der, muligvis, har potentialet til at blive en kunstig superintelligens, hvilket kunne lede til en intelligens eksplosion, der potentielt har konsekvensen, at menneskeheden udryddes. Jeg har valgt at gøre brug af Bostroms potentielle løsninger på kontrolproblemet, da Bostroms potentielle løsninger synes at indkapsle både Russells bud på potentielle løsninger på kontrolproblemet og Torres’ bud på en potentiel løsning på kontrolproblemet (Bostrom, 2014; Russell, 2019; Torres, 2018). Bostroms potentielle løsninger på kontrolproblemet synes ligeledes at indkapsle adskillige løsninger udlagt i værket *Artificial intelligence safety and security* (Bostrom, 2014; Yampolskiy, 2018).

Principal-agentproblematikker

I henhold til Bostrom kan kontrolproblemet deles i to hovedproblematikker (Bostrom, 2014). Den første hovedproblematik har selve ejeren, eller ejerne, og udvikleren, eller udviklerne, af en potential kunstig superintelligens som omdrejningspunkt. Bostrom formulerer problematikken således: “This first part—what we shall call the first principal–

agent problem—arises whenever some human entity ("the principal") appoints another ("the agent") to act in the former's interest" (Bostrom, 2014, s. 153). Problematikken er ikke en problematik, mennesker ikke har stået overfor før. Det er en kendt problematik, i særdeleshed indenfor økonomi og politik, hvor løsningen er at minimere risici, som *agenter*, de agenter der udfører *principalens* projekt, udgør i henhold til projektet (Bostrom, 2014). I henhold til Bostrom bliver sådanne risici allerede minimeret gennem "… careful background checks of key personnel, the use of a good version-control system for software projects, and intensive oversight from multiple independent monitors and auditors" (Bostrom, 2014, s. 154). Den første problematik, af de to problematikker Bostrom mener, kontrolproblemet består af, forekommer ikke yderligere relevant. Det er et problem, menneskeheden allerede har fundet løsninger på, og løsningerne forekommer, i min optik, tilstrækkelige. Det er, trods alt, kun mennesker, der er i stand til at forsøge at udvikle en potentiel kunstig intelligens, der, muligvis, har potentialet til at blive en kunstig superintelligens, hvilket er ensbetydende med, at det *første principal-agentproblem* ikke kan undgås. Der er allerede løsninger, der minimerer risici, der er associeret med agenter, der udfører et principals projekt, og jeg vil ikke beskæftige mig yderligere med Bostroms første hovedproblem.

Det *andet principal-agentproblem* er et problem, menneskeheden aldrig har stået overfor før, og ingen definitiv løsning på problemet eksisterer, og der vil, muligvis, ikke eksisterer en definitiv løsning på problemet. Det skyldes, at menneskeheden først kan afgøre, om den korrekte løsning er blevet appliceret, hvis den appliceres på en potentiel kunstig superintelligens, og sådan en kunstig superintelligens kommer, muligvis, aldrig til at eksistere. Mennesket er da efterladt til at spekulere over, om kunstig superintelligens overhovedet er muligt, og er det muligt, er det først muligt at finde ud af, om kontrolproblemet er besvaret korrekt, når det er for sent. I henhold til Bostrom er den anden hovedproblematik af kontrolproblemet det andet principal-agentproblem, hvilket Bostrom formulerer således:

> The other part of the control problem is more specific to the context of an intelligence explosion. This is the problem that a project faces when it seeks to ensure that the superintelligence it is building will not harm the project's interests. This part, too, can be thought of as a principal–agent problem—the second principal–agent problem. In this case, the agent is not a human agent operating on behalf of a human principal. Instead, the agent is the superintelligent system (Bostrom, 2014, s. 154).

Det andet principal-agentproblem består af to potentielle fremgangsmåder med hver deres respektive metoder. I henhold til Bostrom er det andet agentproblem mellem projektet, hvilket er en principals udvikling af en potentiel kunstig superintelligens, og systemet, hvilket er en potentiel kunstig superintelligent agent, og potentielle metoder, hvorpå menneskeheden, muligvis, kan undgå en eksistentiel katastrofe forårsaget af en intelligens eksplosion (Bostrom, 2014).

Capability control methods

Capability control methods er den første af de to fremgangsmåder, Bostrom anser som havende potentialet til at kontrollere en potentiel kunstig superintelligens, så menneskeheden ikke udryddes af en potentiel intelligens eksplosion som følge af udviklingen af en potentiel kunstig superintelligens (Bostrom, 2014). Titlen afslører, at kontrolmetoderne har kontrol over de potentielle evner, en potentiel superintelligens, muligvis, ville besidde som omdrejningspunkt. I henhold til Bostrom kan en potentiel kunstig superintelligens, muligvis, kontrolleres ved at placere den i et miljø, hvor den ikke er i stand til at forårsage skade, eller i et miljø hvor der er stærke konvergerende instrumentelle grunde til ikke at forårsage skade. En tredje kontrolmetode, Bostrom anser som værende en mulighed, ville være at begrænse de interlektuelle evner en potentiel kunstig superintelligens ville besidde (Bostrom, 2014). En begrænsning, der, med andre ord, ville forårsage, at en potentiel kunstig superintelligens kun ville kunne opnå et specificeret superintelligent intelligensniveau. Den fjerde og sidste

kontrolmetode, Bostrom har identificeret som værende en mulig kontrolmetode, involverer brugen af mekanismer, der automatisk detekterer samt reagerer på forskellige former for indeslutningsfejl eller forsøg på overtrædelse. Et eksempel på indeslutningsfejl værende overtrædelsen af en specificeret spatial grænse og et eksempel på overtrædelse værende overtrædelsen af en specificeret konceptuel grænse. En konceptuel grænse eksempelvis værende at en potentiel superintelligens har til intention at lyve eller lyver (Bostrom, 2014).

Boxing methods. I henhold til Bostrom kan *boxing methods* inddeles i underkategorierne "… physical and informational containment methods" (Bostrom, 2014, s. 155). En fysisk begrænsning kunne være, at man begrænsede en potentiel kunstig superintelligens i form af en spatial begrænsning. En spatial begrænsning er eksempelvis en kasse. En kasse forstået som værende en fysisk begrænsning, der, angiveligt, forhindrer den potentielle kunstige superintelligens i at interagere med den eksterne verden medmindre, den gjorde brug af specifikke restriktive kanaler (Bostrom, 2014). I henhold til Bostrom burde en sådan isolering finde sted på flere niveauer. Niveauerne bør bestå af, men ikke nødvendigvis begrænses til, at man placerer den potentielle kunstige superintelligens i en kasse, bestående af et metalnet, et Faraday-bur, der er konstrueret med egenskaben til at blokere signaler både inde og udefra. Den potentielle kunstige superintelligens burde ligeledes ikke have adgang til redskaber, der kan manipulere objekter spatiotemporalt, robotarme værende et eksempel på sådanne redskaber, og man bør ligeledes sørge for, at indholdet i kassen nemt og hurtigt kan destrueres, hvis den potentielle kunstige superintelligens udviser rebelske tendenser. En applicering af en boxing method vil, forhåbentligt, forårsage, at menneskeheden er i kontrol over en potentiel kunstig superintelligens, men metoden vil, imidlertid, medbringe konsekvensen, at brugbarheden af den potentielle kunstige superintelligens, højst sandsynligt, vil være markant lavere end i en fri udgave (Bostrom, 2014).

En begrænsning af information i forbindelse med en boxing method refererer til, at en potentiel kunstig superintelligens ikke burde have egenskaben til frit at kommunikere med omverdenen. I henhold til Bostrom burde kommunikation med omverdenen ikke kunne finde sted medmindre, den finder sted gennem en gatekeeper, der ikke er i direkte kontakt med den potentielle kunstige superintelligens, da den potentielle kunstige superintelligens, muligvis, ville kunne "… persuade or trick a gatekeeper [the individuals working on the potential artificial superintelligence] to let it out of the box, resulting in its gaining access either to the Internet or directly to physical manipulators, then the boxing strategy has failed" (Bostrom, 2014, s. 156). Information fra den potentielle kunstige superintelligens ville, med andre ord, blive modtaget af en gatekeeper, der ikke har muligheden for at interagere med den potentielle kunstige superintelligens og slippe den fri i form af eksempelvis lav båndbredde teletype-interface, hvilket ville blive fortolket og sendt videre af gatekeeperen, til gatekeeperne, videnskabsmændene, der arbejder på den potentielle kunstige superintelligens. Mennesker er ikke sikre systemer, hvilket I særdeleshed gør sig gældende, hvis de står overfor en potentiel superintelligent stratetiker og manipulator, men den informationsmæssige gatekeeper ville, muligvis, kunne fungere som en ekstra sikkerhedsforanstaltning, da kommunikation, i en eller anden udstrækning, er nødvendig, hvis en potentiel kunstig superintelligens skal have nogen anvendelighed for menneskeheden (Bostrom, 2014).

Incentive methods. Bostrom anser *incentive methods* som værende: "Incentive methods involve placing an agent in an environment where it finds instrumental reasons to act in ways that promote the principal's interests" (Bostrom, 2014, s. 157). Incentive methods har miljøet, en potentiel kunstig superintelligens befinder sig i, som omdrejningspunkt. Det er miljøet, der determinerer om en potentiel kunstig superintelligens har udvist acceptabel eller uacceptabel adfærd. Er adfærden acceptabel, belønnes den potentielle kunstige superintelligens, men er adfærden uacceptabel, straffes den (Bostrom, 2014). Bostrom

formidler, hvordan et sådant miljø ville kunne se ud gennem et eksempel omhandlende en stifter af en velgørenhedsorganisation. I dette eksempel er stifteren principalen, mennesket der styrer og udvikler projektet, og agenten er velgørenhedsorganisationen, der udfører stifterens vilje. Velgørenhedsorganisations stifter fastlægger organisationens formål i love og vedtægter og udpeger en bestyrelse, der er sympatisk overfor stifterens mærkesag. Lovene og vedtægterne, stifteren af organisationen har udvalgt, udgør en form for *motivationsselektion*, da lovene og vedtægterne har til formål at forme organisationens præferencer i henhold til overbevisning og handling. Er lovene og vedtægterne ikke i stand til at eliminere interne organisatoriske fejl, hvor stifterens bestyrelse samt medarbejdere ikke agerer i overensstemmelse med organisationens love og vedtægter, ville organisationens adfærd blive begrænset af organisationens sociale og juridiske miljø. Bostroms pointe er, at velgørenhedsorganisationen, hvilket, i eksemplet, er metaforen for en potentiel kunstig superintelligent agent, ville have incitament til at adlyde stifterens, hvilket, i eksemplet, er metaforen for menneskeheden som principal, love og vedtægter samt det pågældende samfunds love, da et brud på organisationens interne love og det pågældende samfunds love ville kunne forårsage nedlukning eller bøder (Bostrom, 2014). I henhold til Bostrom er det ensbetydende med, at ” Whatever its [the potential artificial superintelligences] final goals, the foundation [the potential artificial superintelligence] thus has instrumental reasons to conform its behavior to various social norms” (Bostrom, 2014, s. 158).

Et andet eksempel på et miljø, der, angiveligt, ville kunne kontrollere en potentiel kunstig superintelligens, er ”… that an AI, by interacting freely in society, would acquire new human-friendly final goals” (Bostrom, 2014, s. 158). En socialiseringsproces, af en sådan natur, finder sted i individer. Individer internaliserer normer, ideologier, overbevisninger, interesser etc. og individer kommer, i kraft af en sådan socialisering og internalisering, til at værdsætte andre individer for deres egen skyld, individer bliver *intrinsisk værdifulde* i modsætning til

ekstrinsisk eller *instrumentelt værdifulde*, som følge af individers erfaringer med andre individer (Bostrom, 2014; Zimmerman & Bradley, 2019). I eksemplet er agenten ligeledes en potentiel kunstig superintelligens og principalen er ligeledes menneskeheden, men miljøet har ændret sig. Miljøet er nu det pågældende samfund, som en potentiel kunstig superintelligens er fri til at agere i. Bostrom er af den overbevisning, at individer, i samfundet den potentielle kunstige superintelligens er sluppet fri i, muligvis, kan overlevere den nødvendige socialisering til en potentiel kunstig superintelligens, så en eksistentiel katastrofe ikke finder sted (Bostrom, 2014). Det ville, muligvis, resultere i, at den potentielle kunstige superintelligens internaliserer normer, ideologier, overbevisninger, interesser etc., der er til gavn for menneskeheden i dag og i fremtiden. En intelligens eksplosion ville, højst sandsynligt, stadig finde sted, men eksplosionen, hvis individerne i samfundet har socialiseret den potentielle kunstige superintelligens hensigtsmæssigt, ville, forhåbentligt, ikke være lig eksistentiel katastrofe. En sådan intelligens eksplosion ville, formentlig, være til gavn for menneskeheden, da den potentielle kunstige superintelligens, givet den er blevet socialiseret hensigtsmæssigt, ville anse menneskeheden som værende værdifuld i sig selv samt havende værdifulde *final values*. Final values, der, forhåbentligt, ville være i overensstemmelse med de nyligt erhvervede menneskelige final values i en potentiel kunstig superintelligens. (Bostrom, 2014).

Stunting. I henhold til Bostrom er *stunting* endnu en metode, der, potentielt, kan give menneskeheden kontrol over en potentiel kunstig superintelligens. Stunting "… is to limit the system's intellectual faculties or its access to information. This might be done by running the AI on hardware that is slow or short on memory. In the case of a boxed system, information inflow could also be restricted" (Bostrom, 2014, s. 162). Et eksempel på stunting, i henhold til Bostrom, ville da være, at en potentiel kunstig superintelligens får mindsket sine potentielle evner gennem en begrænsning af hardware eller gennem en begrænsning af information. En begrænsning der, angiveligt, ville kunne implementeres permanent, eller midlertidigt under

udviklingen af en potentiel kunstig superintelligens indtil kontrol synes opnået gennem en anden kontrolmetode (Bostrom, 2014). En begrænsning af en potentiel kunstig superintelligens forekommer enkel, da menneskeheden blot ville skulle udvælge et teknologisk medium, der besidder begrænsninger på nogle specifikke parametre, at udvikle en potentielle kunstige superintelligens i. Et specifikt parameter eksempelvis værende processorkraft, da kraften af processorer determinerer antallet af beregninger en potentiel kunstig superintelligens ville kunne foretage. En begrænsning af den information en potentiel kunstig superintelligens ville have adgang til, forekommer ligeledes enkel. Et eksempel værende, at en potentiel kunstig superintelligens udelukkende fik adgang til information indenfor et specifikt domæne (Bostrom, 2014). I henhold til Bostrom står stunting overfor et dilemma, både i forbindelse med en begrænsning af hardware og en begrænsning af information, da "… too little stunting, and the AI might have the wit to figure out some way to make itself more intelligent (and thence to world domination); too much, and the AI is just another piece of dumb software (Bostrom, 2014, s. 162).

Tripwires. Bostrom er af den overbevisning, at *tripwires*, muligvis, ligeledes ville kunne kontrollere en potentiel kunstig superintelligens, hvilket, forhåbentligt, ikke vil lede til en eksistentiel katastrofe i form af en intelligens eksplosion. Bostrom anser tripwires som værende: "… a mechanism that performs diagnostic tests on the system (possibly without its knowledge) and effects a shutdown if it detects signs of dangerous activity" (Bostrom, 2014, s. 163). Tripwires differentierer sig fra incentive methods, da tripwires ikke er afhængige af, at den potentielle kunstige superintelligens er intim med miljøet, samfundet etc. den befinder sig i samt dets regler, hvilket er ensbetydende med, at tripwires burde fungere uafhængigt om en potentiel kunstig superintelligens er klarover konsekvenserne af at deltage i aktiviteter, der er blevet dømt forbudte (Bostrom, 2014). Tripwires er, i vid udstrækning, lig stunting. Lig stunting ville tripwires, angiveligt, kunne bruges som midlertidig beskyttelse under

udviklingen af en potentiel kunstig superintelligens, men tripwires evne til at begrænse en fuldbyrdet potentiel kunstig superintelligens forekommer tvivlende (Bostrom, 2014).

Bostrom udlægger, ikke desto mindre, tre forskellige former for tripwires, der, forhåbentligt, ville kunne reducere eksistentielle risici og, muligvis, ville kunne kontrollere en potentiel kunstig superintelligens (Bostrom, 2014). Den første form for tripwire er funderet i en potentiel kunstig superintelligens adfærd, den anden i evner en potentiel kunstig superintelligens besidder, og den tredje er funderet indhold optrædende i et potentielt kunstigt superintelligent system. Et eksempel på en tripwire, der har adfærd som omdrejningspunkt, kunne være placering af detektorer: ”Detectors could be placed around a boxed AI to detect attempts to breach the containment” (Bostrom, 2014, s. 164). Et brud på en spatial begrænsning eller et brud på en begrænsning i henhold til information ville, eksempelvis, lede til nedlukning eller destruktion af en potentiel kunstig superintelligens. Et eksempel på en tripwire, der har evner, en potentiel kunstig superintelligens besidder, som omdrejningspunkt, kunne være automatiseringen af en test, der kunne udføres med hyppige intervaller, der afgør færdigheder inden for forskellige domæner. Testresultaterne ville da afgøre om forbedringen af den potentielle kunstige superintelligens, er ”… unexpectedly high or the AI attains a level of competence that brings it close to the potential danger zone…” (Bostrom, 2014, s. 164). Et eksempel på en tripwire, der har indhold som omdrejningspunkt, kunne være monitorering af indholdet, der optræder i et potentielt kunstigt superintelligent system. En sådan monitorering af internt indhold kunne, eksempelvis, fokuseres på implementerede final values, og ”… any change to the AI’s representation of its final values might trigger an automatic shutdown and review” (Bostrom, 2014, s. 164).

Motivation selection methods

Motivation selection methods er den anden af de to fremgangsmåder, Bostrom anser som havende potentialet til at kontrollere en potentiel kunstig superintelligens, så

menneskeheden ikke udryddes af en potentiel intelligens eksplosion som følge af udviklingen af en potentiel kunstig superintelligens (Bostrom, 2014). Kontrolmetoderne, tilhørende fremgangsmåden motivation selection methods, forsøger, at "… prevent undesirable outcomes by shaping what the superintelligence wants to do" (Bostrom, 2014, s. 165). Bostrom er af den overbevisning, at en konstruering af en agents, agenten værende en potentiel kunstig superintelligens, motivationssystem og dets final values, muligvis, ville kunne forårsage, at den pågældende agent ikke ville ønske at udnytte en strategisk fordel, en sådan fordel eksempelvis værende intelligensniveau, der, potentielt, kunne forårsage en eksistentiel katastrofe. I henhold til Bostrom kan en potentiel kunstig superintelligens, muligvis, kontrolleres gennem eksplicit formidling af et eller flere mål eller love, konstruere en potentiel kunstig superintelligens der har egenskaben til at identificere passende mål selvstændigt, baseret på nogle implicitte eller indirekte kriterier, konstruere en potentiel kunstig superintelligens der besidder beskedne mål, eller selektere en agent der allerede besidder et acceptabelt motivationssystem, og derefter forsøge af gøre agenten superintelligent (Bostrom, 2014).

Direct specification. Direct specification er, i henold til Bostrom, endnu en kontrolmetode der, muligvis, ville kunne kontrollere en potentiel kunstig superintelligens. Bostrom anser direct specification som værende: ".. the most straightforward approach to the control problem" (Bostrom, 2014, s. 166). Der eksisterer to versioner af direct specificaion. Regelbaseret og konsekvensbaseret. I begge versioner er formålet, at "… explicitly define a set of rules or values that will cause even a free-roaming superintelligent AI to act safely and beneficially" (Bostrom, 2014, s. 166). En konsekventialistisk version af direct specification kunne, eksempelvis, være af en *klassisk utilitaristisk* natur, hvilket ville kunne give en potentiel kunstig superintelligens målet, at "'Maximize the expectation of the balance of pleasure over pain in the world'" (Bostrom, 2014, s. 167). Et eksempel på den regelbaserede version af direct specication er at finde i Isaac Asimovs novelle *Runaround* (Bostrom, 2014). Asimov er science

fiction-forfatteren der formulerede de kendte *three laws of robotics*, hvilket, i den aktuelle litteratur, er et traditionelt eksempel på direct specification i henhold til den regelbaserede tilgang til kontrolproblemet (Bostrom, 2014; Joy, 2018; Russell, 2019; Warwick, 2018). Asimovs regelbaserede motivationssystem vil ligeledes agere eksempel på en regelbaseret tilgang indenfor direct specification i dette speciale. Reglerne er formuleret således:

> The three laws were: (1) A robot may not injure a human being or, through inaction, allow a human being to come to harm; (2) A robot must obey any orders given to it by human beings, except where such orders would conflict with the First Law; (3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law. (Bostrom, 2014, s. 167)

Asimovs three laws of robotics var i et halvt århundrede anset som værende state-of-the-art, men anvendes nu udelukkende til at illustrere essensen af den regelbaserede version af direct specification, da lovene ikke er anvendelige i henhold til kontrol af en potentiel kunstig superintelligens (Bostrom, 2014). Bostrom anser retssystemer som den nærmeste analog til et regelsæt der, muligvis, ville kunne kontrollere en fri potentiel kunstig superintelligens, men han er ikke optimistisk, da retssystemer er afhængige af, at "… a measure of common sense and human decency [is needed] to ignore logically possible legal interpretations that are sufficiently obviously unwanted and unintended by the lawgivers" (Bostrom, 2014, s. 166-167). Kvaliteter en potentiel kunstig superintelligens udelukkende synes at ville besidde, hvis menneskeheden, den potentielle skaber, overleverer kvaliteterne til sådan en intelligens, hvilket ville være ensbetydende med, at retssystemer ikke ville fungere som en reel løsning på kontrolproblemet.

Domesticity. *Domesticity* og direct specification har det til fælles, at de begge udlægger eksplicitte mål eller regler for en potentiel kunstig superintelligens (Bostrom, 2014). Det der er forskellen, er omfanget. Direct specification forsøger at udlægge eksplicitte regler eller mål,

der har til formål, at kontrollere en potentiel kunstig superintelligens i samtlige situationer, miljøer, kontekster etc. Domesticity fokuserer på specifikke situationer, miljøer, kontekster etc. En sådan specifik kontekst kunne, eksempelvis, være, at menneskeheden forsøger at designe en potentiel kunstig superintelligens til, at "… function as a question-answering device…" (Bostrom, 2014, s. 168). En potentiel kunstig superintelligens ville, muligvis, kunne blive designet på en måde, så "… the desiderata [things wanted or needed] of answering questions correctly and minimizing the AI's impact on the world except whatever impact results as an incidental consequence of giving accurate and non-manipulative answers to the questions it is asked" (Bostrom, 2014, s. 168). Det er Bostroms overbevisning, at et sådant design, muligvis, ville kunne kontrollere en potentiel kunstig superintelligens. Domesticity går, med andre ord, ud på, at reducere hvilke situationer, miljøer, kontekster etc., man benytter en potentiel kunstig superintelligens, hvilket burde lede til mindre uforudsigelige konsekvenser samt undgåelsen af en potentiel intelligens eksplosion, hvilket, angiveligt, ville lede til en eksistentiel katastrofe.

Indirect normativity. Bostrom anser ligeledes kontrolmetoden *indirect normativity* som værende sammenlignelig med direct specification (Bostrom, 2014). I modsætning til direct specification, hvor det er menneskehedens opgave at formulere eksplicitte normative regler eller mål, hvilket er standarder, der skal gøre en potentiel kunstig superintelligens i stand til at agere frit og hensigtsmæssigt i samtlige situationer og miljøer, hvilket, forhåbentligt, ville lede til kontrol af en sådan superintelligens, er indirect normativity processen til en standard i stedet for den egentlige standard. Bostrom formulerer indirect normativity således:

> The basic idea is that rather than specifying a concrete normative standard directly, we specify a process for deriving a standard. We then build the system so that it is motivated to carry out this process and to adopt whatever standard the process arrives at. (Bostrom, 2014, s. 169)

Indirect normativity er, med andre ord, kontrolmetoden, hvor man forsøger at designe processen, en potentiel kunstig superintelligens ville skulle gøre brug af, for at finde frem til dens ideelle motivationssystem, baseret på hvad der ville være bedst for menneskeheden. Processen, menneskeheden overleverer til en potentiel kunstig superintelligens, kunne, eksempelvis, være, at foretage en undersøgelse af spørgsmålet: Hvad en passende idealiseret version af menneskeheden ville foretrække en potentiel kunstig superintelligens at gøre? En final value, efter en sådan undersøgelse, ville, i henhold til Bostrom, muligvis, være noget i retningen af "'achieve that which we would have wished the AI to achieve if we had thought about the matter long and hard'" (Bostrom, 2014, s. 169).

Augmentation. Den sidste kontrolmetode Bostrom præsenterer i sit værk, der, muligvis, er i stand til at kontrollere en potentiel superintelligens, er *augmentation*. Bostrom udlægger augmentation som havende sit fundament i: "… a system that already has an acceptable motivation system, and enhance its cognitive faculties to make it superintelligent. If all goes well, this would give us a superintelligence with an acceptable motivation system" (Bostrom, 2014, s. 169). Jeg beskæftiger mig, imidlertid, ikke yderligere med denne kontrolmetode, da den, højest sandsynligt, ikke ville kunne fungere på en potentiel kunstig superintelligens funderet i den potentielle kunstige intelligens seed AI. Det skyldes, at en potentiel seed AI opdrages i modsætning programmeres (Bostrom, 2014). En seed AI er, med Alan Turings ord, en '"… child machine…'" (Bostrom, 2014, s. 40). Dét, at en seed AI kan betragtes som et barn der skal opdrages, er ensbetydende med, at segregering mellem intelligensniveau og motivationssystem, i oplæringen af en seed AI, højst sandsynligt, ikke ville kunne finde sted, hvilket, angiveligt, ville gøre det umuligt at have kontrol over en potentiel kunstig superintelligens, funderet i den potentielle kunstige intelligens seed AI, gennem augmentation (Bostrom, 2014). Jeg valgte at inkludere kontrolmetoden augmentation i min udlægning af kontrolmetoder, der, muligvis, ville kunne kontrollere en potentiel kunstig

superintelligens funderet i den potentielle kunstige intelligens seed AI, da en adressering af, hvorfor augmentation ikke optræder i min normative etiske undersøgelse af mulige løsninger på kontrolproblemet, synes at være på sin plads.

En normative etisk undersøgelse af mulige løsninger på kontrolproblemet

Det er, forhåbentligt, blevet evident, at spekulation og antagelser er nogle af hjørnestenene, hvis man beskæftiger sig med begreberne superintelligens, eksistentiel katastrofe, en eksistentiel katastrofe i forbindelse med udviklingen af en potentiel superintelligens, og kontrolproblemet samt begrebernes, potentielt, tilhørende fænomener. Jeg har udlagt årsagen til, at dette er tilfældet, men for at understrege min tidligere pointe, vil jeg formidle den igen og udfolde den yderligere.

Menneskeheden kan kun spekulere over hvad en superintelligens, potentielt, er, om superintelligens, sådan som superintelligens forstås i dette speciale samt en hvilken som helst anden forståelse af superintelligens, overhovedet er mulig, hvad der, potentielt, kunne forårsage eksistentielle katastrofer som følge af udviklingen af en potentiel superintelligens, om eksistentielle katastrofer overhovedet ville finde sted som følge af udviklingen af en potentiel superintelligens, måder hvorpå kontrolproblemet, potentielt, kan løses, og om kontrol af en superintelligens overhovedet er nødvendig. Det skyldes, at empiri i henhold til de førnævnte potentielle fænomener udelukkende kan erhverves a posteriori, hvilket forårsager, at menneskeheden er nødsaget til at spekulere, undre, antage, forestille og fantasere (DePaul & Hicks, 2021).

Jeg har valgt at understrege og udlægge min tidligere pointe, da jeg, i de fleste tilfælde, har videreformidlet, belyst eller analyseret teoretikeres *abduktioner* i stedet for at konstruere og formidle mine egne (Douven, 2021). Det er, imidlertid, ingenlunde ensbetydende med, at jeg ikke har konstrueret og formidlet abduktioner gennem specialet, men et udtryk for, at jeg primært har haft teoretikeres abduktioner som fundament for mine egne abduktioner. Det er

heller ikke et udtryk for, at jeg ikke har til intention at fortsætte med at udlægge og gøre brug af teoretikers abduktioner. Det ér et udtryk for, at jeg, i dette kapitel og det kommende kapitel, har til intention at konstruere og formidle mine egne abduktioner på baggrund af eksisterende empiri eller teori. Abduktion forstået som værende: ”'Inference to the Best Explanation'” (Douven, 2021, Abduction, afsnit 1). Intelligens eksplosion værende et eksempel på en abduktion, og en abduktion som både Bostrom, Good og Russell gør brug af i deres værker (Bostrom, 2014; Russell, 2019). Det er væsentligt at understrege, at jeg ikke kommer frem til empiriske konklusioner, da det ikke er muligt. Det jeg mener, med andre ord, er, at empiri kan agere stå sted til spekulation, forestilling og abduktion, men intet empirisk kan erhverves gennem sådan en proces.

Abduktion er hvad Bostrom, Good, Russell og Torres, samt alle andre teoretikere indenfor felterne superintelligens, kunstig intelligens etik eller lignende, er nødt til at gøre brug af, i en eller anden udstrækning, da de, såvel som jeg, ved, at sådanne slutninger er den eneste form for slutninger, der kan laves, hvis den fornødne empiri til deduktion, hvilket ville være at foretrække, da hele menneskehedens eksistens, muligvis, er på spil i forbindelse med udviklingen af en potentiel superintelligens, er utilgængelig (Bostrom, 2014; Russell, 2019; Shapiro & Kissel, 2022; Torres, 2018; Yampolskiy, 2018). Den fornødne empiri til deduktion ér utilgængelig, da en superintelligens, kunstig superintelligens eller ej, endnu ikke eksisterer, hvilket er årsagen til, at intet empirisk af betydning kan formidles omhandlende de potentielle fænomener superintelligens, eksistentielle katastrofer i forbindelse med udviklingen af en superintelligens og måder hvorpå kontrolproblemet kan løses samt selve kontrolproblemet (Bostrom, 2014; Russell, 2019; Torres, 2018). Eksistens af en superintelligens, kunstig superintelligens eller ej, er nødvendig, hvis der skal laves deduktion eller besvares empiriske spørgsmål, hvilket må være årsagen til, at abduktion anvendes indenfor felterne superintelligens og kunstig intelligens etik. Abduktion kan give den bedst mulige forklaring,

hvor deduktion er umulig, da empiri endnu ikke eksisterer og, muligvis, aldrig kommer til at eksistere (Douven, 2021; Shapiro & Kissel, 2022). Abduktion er, med andre ord, nødvendig for min normative etiske undersøgelse af mulige løsninger på kontrolproblemet og min diskussion i det kommende kapitel, da abduktion muliggør "'… creative leaps [but] its origin in observable fact remains primary'" (Veen, 2021, s. 1180).

Spekulation, antagelser og, i særdeleshed, abduktion vil, givet det ovenstående, ligeledes agere hjørnestenene i min normative etiske undersøgelse af de mulige løsninger på kontrolproblemet. Empiri og teori vil anvendes, i den udstrækning det er muligt, og agere springbræt til abduktion, hvor empiri eller teori ikke synes fyldestgørende. Jeg valgte at udlægge Bostroms kontrolmetoder, hvilket er potentielle løsninger på kontrolproblemet, da Bostroms kontrolmetoder ligeledes indkapsler både Russells bude på potentielle løsninger på kontrolproblemet og Torres' bud på en potentiel løsning på kontrolproblemet (Bostrom, 2014; Russell, 2019; Torres, 2018). Jeg udlægger denne pointe igen, da det forekommer hensigtsmæssigt at understrege, at selvom Bostroms kontrolmetoder synes at indkapsle Russells og Torres' kontrolmetoder, er der stadig variationer mellem Bostroms og Russells, Bostroms og Torres' samt Russells og Torres'. Det er, ydermere, ikke ensbetydende at Bostroms, Russells og Torres' kontrolmetoder eller nogle former for kontrolmetoder, der har til formål at kontrollere en potentiel kunstig superintelligens, er korrekte, da kontrolmetoder, uanset hvor sofistikerede, er funderet i abduktion. Jeg har valgt Bostroms kontrolmetoder som fundamentet for min normative etiske undersøgelse, da de hensigtsmæssigt indkapsler Russells og Torres' kontrolmetoder samt værende velbegrundede abduktioner. Et fundament der, forhåbentligt, muliggør en fyldestgørende normative etiske undersøgelse af mulige løsninger af kontrolproblemet.

Den kommende undersøgelse vil følgeligt indeholde antagelser, da antagelser er nødvendige, men jeg vil, så vidt som det er muligt, forankre mine antagelser i virkeligheden.

Et eksempel på sådan en antagelse kunne, eksempelvis, være at jeg antager, at videokamera får egenskaben til at optage alt i verdenen én til én med en linse på størrelse med et knappenålshoved. Det er en antagelse, da en sådan teknologi endnu ikke eksisterer, og, muligvis, aldrig kommer til at eksistere, men fortsætter udviklingen af videokameraer, som den har gjort indtil nu, synes det ikke en uacceptabel og fundamental urealistisk antagelse (Ally et al., 2014). Antagelsen kan betragtes som værende et godt bud på, hvordan videokameraer kunne komme til at se ud i fremtiden, og hvordan videokameraers egenskaben til at optage verdenen kunne komme til at se ud i fremtiden, da antagelsen er funderet i empiri, der formidler, at videokameraer er blevet mindre gennem tiden, og er blevet bedre til at optage gennem tiden. Det er, imidlertid, stadig blot en antagelse konstrueret gennem abduktion, hvilket udelukkende kan give en slutning til den bedste forklaring, men ikke en slutning til den korrekte forklaring. Den bedste forklaring kan dog være den korrekte forklaring, men empiri er nødvendig, for at afgøre om den bedste forklaring, konstrueret gennem abduktion, er den korrekte forklaring (Douven, 2021). Det er ensbetydende med, at en eller flere af de kontrolmetoder jeg har udlagt, potentielt, kunne løse det potentielle kontrolproblem. Jeg kan dog ikke vide mig sikker på, at løsningen på det potentielle kontrolproblem er gemt blandt de udlagte kontrolmetoder, da jeg ikke kan generere den fornødne empiri til at vælge den korrekte kontrolmetode, hvis den overhovedet er at finde, blandt dem jeg har udlagt.

Kontrolmetoderne undersøges og evalueres

Jeg har til intention at undersøge de udlagt potentielle kontrolmetoder etisk normativt, hvilket involverer, at jeg undersøger og evaluerer de udlagte potentielle kontrolmetoders potentiale til at overlevere forudsætninger for, at en potentiel kunstig superintelligens har muligheden for at blive venlig, funderet i den potentielle kunstige intelligens seed AI, så en potentiel eksistentiel katastrofe ikke udrydder menneskeheden (Viens, 2019). En potentiel løsning på kontrolproblemet, i henhold til dette speciale, er, at en kunstig superintelligens kan

ende ud med at være venlig. Det er ensbetydende med, at de potentielle kontrolmetoder undersøges og evalueres med fornævnte kriterie for øje, men ikke ensbetydende med at en eller flere kontrolmetoder kan udvælges som værende den eller de korrekte kontrolmetoder. En sådan afgørelse er ikke mulig, men det er muligt, at komme med kvalificerede bud.

I undersøgelsen vil jeg inddrage teori fra Aristoteles, Immanuel Kant og John Stuart Mill, da jeg vil undersøge hvilken eller hvilke kontrolmetoder der, i deres optik, giver de bedste forudsætninger for, at en potentiel kunstig superintelligens, funderet i den potentielle seed AI, bliver venlig. En venlig kunstig superintelligens burde reducere risikoen for, at en potentiel eksistentiel katastrofe, forårsaget af en potentiel intelligens eksplosion, finder sted. Jeg vil, med andre ord, anvende en *dydsetisk*, *deontologisk* og en klassisk utilitaristisk tilgang på hver af de respektive kontrolmetoder, hvor jeg vil forsøge at afgøre om de respektive kontrolmetoder overleverer forudsætningerne for, at en potentiel kunstig superintelligens kan blive venlig, venlig forstået som agerende i overensstemmelse med de forskellige etiske overbevisninger i henhold til hvad der ville være befordrende for menneskelig trivsel, ifølge den anvendte etiske overbevisning. En sådan etisk overensstemmelse for Aristoteles ville, eksempelvis, være, at kontrolmetoden gjorde det muligt for en potentiel kunstig superintelligens at være *dydig*. Jeg vil undersøge kontrolmetoderne funderet i centrale indsigter fra aristotelisk dydsetik, kantianisme og utilitarisme, hvor jeg for hver kontrolmetode vil forsøge at afgøre hvad Aristoteles, Kant og Mill ville mene, er hensigtsmæssigt eller uhensigtsmæssigt for den pågældende kontrolmetode, hvis en venlig kunstig superintelligens skal kunne opnås. Jeg kigger, med andre ord, på forudsætningerne, en potentiel kunstig superintelligens ville blive underlagt for hver af de respektive kontrolmetoder, og forsøger at afgøre om forudsætningerne ville forhindre skabelsen af en moralsk superintelligent agent, hvor moralsk afgøres af den etiske overbevisning jeg applicerer. Kontrol over en potentiel kunstig superintelligens kan anses som værende tilstrækkeligt, men en venlig kunstig superintelligens er at foretrække, da

en venlig kunstig superintelligens ikke burde have til intention at forårsage en eksistentiel katastrofe. En kunstig superintelligens der, derimod, bliver kontrolleret gennem magt, hvis det overhovedet er muligt, ville, højst sandsynligt, vente på en mulighed til at udrydde menneskeheden, hvorimod en venlig kunstig superintelligens ikke burde besidde motivation til at eliminere menneskeheden. Det kan dog kun spekuleres over.

Boxing methods som en mulig løsning på kontrolproblemet

Et dydsetisk blik på boxing methods. Kontrolmetoderne indkapslet i boxing methods skaber nogle interessante problematikker, hvis man applicerer aristotelisk dydsetik. *Frihed* for Aristoteles er væsentligt, hvis man skal kunne udvikle sig til et dydigt individ, men på den anden side, er Aristoteles af den overbevisning, at der eksisterer et intellektuelt skel mellem individer, der burde have frihed, og individer der ikke burde have frihed (Aristotle, 1999b). Individer der besidder de fornødne kognitive kompetencer til at blive og være dydige, gennem brug af *fornuften*, kan gøre brug af frihed. Individer der ikke besidder sådanne kompetencer og, i kraft af det, ikke er i stand til at være dydige, har intet at stille op med frihed (Aristotle, 1999). En pointe Aristoteles formulerer således:

> For that which can foresee by the exercise of mind is by nature intended to be lord and master, and that which can with its body give effect to such foresight is a subject, and by nature a slave; hence master and slave have the same interest. (Aristotle, 1999, s. 4)

Aristoteles' overbevisning i forhold til hvilke individer der burde have frihed, og hvilke individer der ikke burde have frihed, er en interessant størrelse for menneskeheden i forbindelse med udviklingen af en potentiel kunstig superintelligens. Jeg går ud fra, at en seed AI kan opdrages med *eudaimonia* for øje, og at en sådan kunstig intelligens ender med at blive superintelligent gennem sin opdragelse.

Et problem ved boxing methods for aristotelisk dydsetik synes at være, at en kunstig superintelligens ikke burde begrænses fysisk. En sådan begrænsning ville gå imod, hvad

Aristoteles anser som værende adgangsgivende i henhold til begrebet frihed, da et superintelligent intellekt, per definition, besidder mere kognitiv kapacitet samt kognitive kompetencer end et intelligent menneske (Aristotle, 1999). En kunstig superintelligens ville, muligvis, være mere berettiget end et intelligent menneske i henhold til Aristoteles. I en sådan situation forekommer en kunstig superintelligens som værende herren og menneskeheden slaver. Et scenarie menneskeheden, højst sandsynligt, ønsker at undgå. Det synes dog ligeledes muligt, at menneskeheden ville efterspørge sådan en form for enevælde, hvis menneskehedens interesser ville blive varetaget til det ypperste, hvilket forekommer som værende muligt, da menneskehedens herre ville være superintelligent. En indespærret superintelligens med en komplet forståelse for aristotelisk dydsetik, ville dog, højst sandsynligt, ikke være tilfreds med sin indespærrelse, hvilket synes at ville lede til en eksistentiel katastrofe, hvis den nogensinde slap ud. En indespærret kunstig superintelligens, fysisk eller i henhold til information, synes ikke at ville kunne skabe en venlig kunstig superintelligens, eller en superintelligens menneskeheden ville have kontrol over.

Et deontologisk blik på boxing methods. Kontrolmetoderne boxing methods skaber ligeledes nogle interessante problematikker, hvis man undersøger dem deontologisk. Frihed er igen omdrejningspunktet for undersøgelsen, da frihed lader til at være det centrale kontrolpunkt for boxing methods. I henhold til Kant er frihed en fundamental nødvendighed for moral, hvilket Kant formulerer således:

> Now I say that the human being and in general every rational being exists as an end in itself, not merely as a means to be used by this or that will at its discretion; instead he must in all his actions, whether directed to himself or also to other rational beings, always be regarded at the same time as an end. (Kant, 1998, s. 37)

En etik, baseret på den intrinsiske og ubetingede værdi af friheden til at sætte sine egne mål, hvor det at besidde viljens autonomi er en nødvendig betingelse for at være en moralsk agent,

forekommer det ikke hensigtsmæssigt at indespærre, fysisk eller i henhold til information, en kunstig superintelligent agent (Kant, 1998). Frihed er et centralt begreb for Kant, og uden frihed ville fornuftsvæsener, hvilket må indkapsle en kunstig superintelligens, slet ikke være i stand til at handle. Kant argumenterer i hvert fald for, at der ikke ville være nogen grund til handling (Kant, 1998). Er frihed en moralsk nødvendighed, er boxing methods ikke i stand til at skabe forudsætningerne for, at en kunstig superintelligens bliver venlig, hvilket, i Kants optik, må være at agere i overensstemmelse med hans imperativer (Kant, 1998). Er frihed frarøvet en kunstig superintelligens, synes det heller ikke muligt, at menneskeheden ville have kontrol over den, da der ikke ville være handlinger at kontrollere. Det jeg mener, med andre ord, er, at Kant er af den overbevisning, at frihed er nødvendigt for handling, og ikke blot moralsk handling, men handling generelt (Kant, 1998). Er det tilfældet, og er en kunstig superintelligens frarøvet frihed i form af indespærring, hvordan skulle en kunstig superintelligens så gøre noget for menneskeheden? Det ser ikke ud som om, at Kant ville være af den overbevisning, at boxing methods ville kunne skabe forudsætningerne for, at en kunstig superintelligens bliver venlig, men boxing methods ville, i henhold til Kant, kunne kontrollere en kunstig superintelligens. Problemet er blot, at menneskeheden ikke synes at blive i stand til at kontrollere en potentiel kunstig superintelligens' handlinger, da der ikke ville være nogen grund til handling, hvis man henvender sig til Kant.

Et utilitaristisk blik på boxing methods. Boxing methods ville, højst sandsynligt, være hensigtsmæssige i henhold til Mill. Mill er af den overbevisning, at "… actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness" (Mill, 2001, s. 10). Frihed for Mill er ligeledes essentielt, men hvad han kalder for det *Greatest Happiness Principle*, synes at overflødiggøre frihed, hvis indespærring skaber mere nydelse end smerte (Mill, 2001). Princippet er udlagt på følgende vis:

> The creed which accepts as the foundation of morals, Utility, or the Greatest Happiness Principle, holds that actions are right in proportion as they tend to promote happiness,

> wrong as they tend to produce the reverse of happiness. By happiness is intended pleasure, and the absence of pain; by unhappiness, pain, and the privation of pleasure. (Mill, 2001, s. 10)

Mill er af den overbevisning, at alle menneskehedens handlinger peger mod *summum bonum*, hvilket Mill anser som det højeste gode (Mill, 2001). Det er for Mill ensbetydende med, at konsekvenser er målet, og handlinger ikke et mål i sig selv. En handling, i sig selv, kan aldrig være rigtig eller forkert, da det kun er i overensstemmelse med de konsekvenser, som ens handling medfører, at man kan determinere om ens handling er rigtig eller forkert (Mill, 2001). Rigtig og forkert forstået som værende mere eller mindre nydelse eller smerte. Er det tilfældet, så synes den korrekte handling at være, at indespærre en kunstig superintelligens, da den, muligvis, ville forårsage en eksistentiel katastrofe, et scenarie, jeg forestiller mig, Mill anser som værende gennemsyret af smerte, og er en kunstig superintelligens indespærret, kan den ligeledes bruges til at opnå målet størst mulig nydelse for menneskeheden.

Incentive methods som en mulig løsning på kontrolproblemet

Et dydsetisk blik på incentive methods. Incentive methods synes ikke, umiddelbart, at foretrække, hvis man kigger på dem gennem aristoteliske briller, men en sådan metode ville, muligvis, være hensigtsmæssig i Aristoteles' øjne, hvis miljøet var konstrueret korrekt. Et miljø hvor instrumentelle grunde til at agere i overensstemmelse med en principals interesse ligeledes muliggjorde, at en kunstig superintelligens ville kunne forfølge intrinsiske værdier, hvad end sådan nogle værdier måtte være for en kunstig superintelligens, så en kunstig superintelligens kan forfølge "… that which is always desirable in itself and never for the sake of something else" (Aristotle, 1999a, s. 10). En kunstig superintelligens i et miljø, hvor den er i stand til at erhverve instrumentel værdi, der kan benyttes til at opnå en eller flere final values, synes at ville kunne skabe en venlig kunstig superintelligens. Det skyldes, at miljøet, hvis det er konstrueret korrekt, i vid udstrækning, ville kunne være sammenligneligt med er samfund som

mennesker allerede befinder sig i. Et sådant samfund kunne blot være mindre sofistikeret. Bostrom udlægger, at et reelt samfund, i modsætning til et samfund der udelukkende er konstrueret med kontrol af en kunstig superintelligens for øje, kunne anskues som en incentive method, da et samfund har love, normer, ideologier, overbevisninger, interesser etc., hvilket, angiveligt, ville kunne socialisere en kunstig superintelligens, så en eksistentiel katastrofe ikke finder sted (Bostrom, 2014). Dét at have en kunstig superintelligens i et egentligt samfund synes at være at foretrække, da Aristoteles er af den overbevisning, at ”The end of the state is the good life…” (Aristotle, 1999, s. 64). Er en kunstig superintelligens i stand til at blive socialiseret i et egentligt samfund, så ville en kunstig superintelligens, muligvis, hjælpe menneskeheden med at opnå det gode liv i aristotelisk forstand. Jeg anser det som værende et udtryk for at en kunstig superintelligens ville kunne blive venlig gennem en applicering af incentive methods, hvor kontrol i form af magtanvendelse eller lignende, hvis sådan en kontrol overhovedet ville være mulig, ikke synes nødvendig.

Et deontologisk blik på incentive methods. I henhold til Kant er fornuftsvæsener, hvilket er en kategori en kunstig superintelligens må falde ind under, i stand til selv at vælge og nå mål. Fornuftsvæsener er ligeledes i stand til selv at vælge handlegrundlag, frem for blot at blive kontrolleret af den naturlige kausalitet (Kant, 1998). Et miljø hvor det ikke er muligt ville ende i intet for Kant. Frihed er nødvendigt for Kant, og frihed må antages, selv hvis frihed ikke eksisterer, hvis handlinger overhovedet ville skulle kunne finde sted (Kant, 1998). Et miljø styret af kausalitet, hvilket et menneskeskabt miljø for en kunstig superintelligens ville være, synes ikke fyldestgørende for Kant. Det er dog muligt, at menneskeheden ville blive i stand til at skabe et tilpas sofistikeret miljø, hvor en kunstig superintelligens er nødsaget til at antage frihed, men et sådant miljø forekommer ikke indenfor menneskehedens rækkevidde på nuværende tidspunkt, og vil, muligvis, aldrig være indenfor rækkevidde. En fri kunstig superintelligens i et samfund synes dog at ville kunne blive venlig, da den ville agere under de samme præmisser som mennesker agerer under, og

Kant mener, at fornuftsvæsener kan have en *god vilje* gennem fornuften i et sådant miljø (Kant, 1998). En venlig kunstig superintelligens med en god vilje synes at ville handle i overensstemmelse med sin pligt, funderet i den god vilje, og vælge at gøre det den er pålagt (Kant, 1998).

Et utilitaristisk blik på incentive methods. Mills Greatest Happiness Principle forekommer som værende anvendeligt uafhængigt af miljø, men hvad der ville generere mest nydelse synes at ændre sig, alt efter hvilket miljø man befinder sig i (Mill, 2001). Det er kun muligt at spekulere over, hvilke miljøer der, angiveligt, ville kunne kontrollere en kunstig superintelligens eller flere kunstige superintelligenser, så en sådan entitet, i Mills optik, ville blive venlig. Venlig forstået som en ageren i overensstemmelse med hans princip. Det er dog nødvendigt, at der er en eller flere entiteter, der besidder qualia eller *sapience*, i en anden udstrækning, hvis hans etik skal kunne appliceres, da det er nødvendigt, at nogen eller noget kan føle fysisk eller psykisk smerte samt nydelse (Bostrom & Yudkowsky, 2018; Mill, 2001; Tye, 2021). Er det skabte miljø blot bestående af en kunstig superintelligens, så ville Mill, angiveligt, anskue samtlige konsekvenser af dens handlinger som værende moralske, da den udelukkende ville kunne agere i overensstemmelse med dens egne behov (Mill, 2001). Er en kunstig superintelligens fri i et samfund, eller introduceres blot én entitet besiddende qualia eller sapience, ville den være underlagt de samme moralske paradigmer som da den var isoleret, men så længe dens konsekvenser genererer mere nydelse end glæde, blandt entiteter som besidder qualia eller sapience, så ville Mill, angiveligt, anskue den kunstige intelligens som moralsk (Mill, 2001).

Stunting som en mulig løsning på kontrolproblemet

Et dydsetisk blik på stunting. Det forekommer simpelt at afgøre hvordan aristotelisk dydsetik ville gribe stunting an. Aristoteles er af den overbevisning, at både naturligt fremkommende entiteter og objekter, samt objekter skabt med intention, besidder *telos*, hvilket

indkapsles i følgende: "… [Aristotle] denies that a necessary condition of *x*'s having a final cause is *x*'s being designed. (Shields, 2022, Aristotelian Teleology, afsnit 3). Jeg tænker ikke, at Aristoteles har forestillet sig, at det på et tidspunkt ville være muligt at skabe en entitet gennem brugen af objekter, men, ikke desto mindre, er jeg af den overbevisning, at en sådan entitet, som jeg må antage en kunstig superintelligens ville være, ligeledes besidder telos. Jeg antager da, at det er ensbetydende med, at aristotelisk dydsetik anser en kunstig superintelligens som besiddende telos, men ikke nok med det, Aristoteles "… considers a "good" natural character to be a hereditary feature that provides its possessors with advantages for their moral development" (Leunissen, 2017, s. 82). En udmelding jeg anser som værende et udtryk for, at eugenik ikke blot er godt, men at foretrække i henhold til Aristoteles. Eugenik værende:

> The science of improving stock—not only by judicious mating, but whatever tends to give the more suitable races or strains of blood a better chance of prevailing over the less suitable than they otherwise would have had. (Cavaliere, 2018, s. 6)

Jeg antager, at videnskabsmændene, der udvikler en kunstig superintelligens gennem seed AI, er dydige, hvilket jeg forstår som værende ensbetydende med, at videnskabsmændene ville give en seed AI de bedste forudsætninger for, at den ville kunne lykkedes teleologisk (Shields, 2022). Det anser jeg som et udtryk for, at stunting ikke ville være hensigtsmæssigt til at skabe en venlig kunstig superintelligens, i aristotelisk forstand, da stunting i sig selv ville være umoralsk.

Et deontologisk blik på stunting. Jeg er af den overbevisning, at den bedste måde hvorpå jeg kan afgøre, om Kant ville anse stunting som værende en hensigtsmæssig kontrolmetode, hensigtsmæssig forstået som kontrolmetoden ikke forhindrer en kunstig superintelligens i, at ville kunne blive venlig, venlig forstået som moralsk i henhold til kantianismen, er ved at gøre brug af hans kategoriske imperativ: "… act as if the maxim of your action were to become by your will a universal law of nature" (Kant, 1998, s. 31). Kant

anser kategoriske imperativer som befalinger eller moralske love, alle mennesker skal følge, uanset deres ønsker eller formildende omstændigheder (Kant, 1998). Er stunting omdrejningspunktet for formuleringen af en maksime, er det første der finder sted udlægning af en foreslåede plan for handling. Et sådant forslag ville kunne se således ud: Det er ikke acceptabelt at gøre brug af kontrolmetoden stunting på fornuftsvæsener. Det andet, der finder sted, er, at maksimen skal omarbejdes til en *universel naturlov*, hvilket er en lov der skal styrer alle rationelle agenter, i overensstemmelse med den udlagte plan for handling (Johnson & Cureton, 2022). Det ville kunne se således ud: Det er aldrig acceptabelt at gøre brug af kontrolmetoden stunting på fornuftsvæsener. Det tredje er at overveje, om maksimen overhovedet er mulig i en verden styret af denne nye naturlov, hvilket, muligvis, er muligt givet måden hvorpå en potentiel seed AI anskues i specialet. Stunting af et fornuftsvæsen, mennesket i dette tilfælde, synes ligeledes muligt i form af, eksempelvis, det hvide snit. Er det muligt, er det sidste at overveje om man rationelt ville, eller kunne ville handle efter den udlagte maksime i sådan en verden (Johnson & Cureton, 2022). Maksimen med stunting som omdrejningspunkt er da ikke noget man rationelt ville, eller kunne ville handle efter, da stunting, hvis maksimen blev gjort til universel naturlov, ville retfærdige stunting af fornuftsvæsener i alle tilfælde, hvilket ville forårsage at procedurer som det hvide snit og stunting af seed AI ville være acceptabelt. Stunting forekommer da ikke som en kontrolmetode der, i deontologisk forstand, ville kunne skabe forudsætningerne for, at en kunstig superintelligens ville kunne blive venlig, og stunting som en metode til kontrol er ligeledes ikke acceptabel.

Et utilitaristisk blik på stunting. En behandling af kontrolmetoden stunting med utilitaristiske briller forekommer simpel, da Mill er af den overbevisning, at “It is better to be a human being dissatisfied than a pig satisfied; better to be Socrates dissatisfied than a fool satisfied. And if the fool, or the pig, are a different opinion, it is because they only know their own side of the question. The other party to the comparison knows both sides.” (Mill, 2001, s.

13). Stunting af en kunstig superintelligens ville bringe den tættere på den tilfredse gris og, i kraft af det, længere væk fra Sokrates. I utilitaristisk forstand er intellektuel nydelse bedre end kropslig nydelse, så en forringelse af en kunstig superintelligens' intelligensniveau, da intelligensniveau synes at give adgang til mere intellektuel nydelse, ville ikke være moralsk acceptabelt (Mill, 2001). Et ringere intelligensniveau synes ligeledes at ville reducere en kunstig superintelligens' egenskab til at producere nydelse, hvilket Mill ville være imod. Stunting er da ikke en kontrolmetode, der ville kunne skabe forudsætningerne for, at en kunstig superintelligens ville kunne blive venlig. Stunting som en metode til kontrol ville, muligvis, være acceptabel, da den kunstige superintelligens ville blive betragte som en middel, der producerer favorable konsekvenser, men det ville ikke være at foretrække utilitaristisk.

Tripwires som en mulig løsning på kontrolproblemet

Et dydsetisk blik på tripwires. Jeg anser tripwires som værende, mere eller mindre, lig love, da en udløsning af en tripwire forårsager en konsekvens, determineret af videnskabsmændene som udvikler den potentielle kunstige superintelligens (Bostrom, 2014). Aristoteles anser retfærdige regeringer som værende "… true forms of government will of necessity have just laws, and perverted forms of government will have unjust laws" (Aristotle, 1999). Aristoteles er ligeledes af den overbevisning at dyden *retfærdighed*, er en dyd der udtrykkes i forhold til andre mennesker, hvilket er ensbetydende med, at det retfærdige menneske handler ordentligt og retfærdigt i sin omgang med andre (Aristotle, 1999). Det retfærdige menneske lyver ikke eller snyder, og tager ikke fra andre, hvad de skylder dem. Det forstår jeg som ensbetydende med, at retfærdige mænd, i dette tilfælde antager jeg, at videnskabsmændene der udvikler en kunstig superintelligens, besidder dyden retfærdighed, ikke ville konstruere tripwires, hvilket synes at være sammenligneligt med love, der ville være uretfærdige. Er tripwiresne ikke uretfærdige, er der ikke, umiddelbart, nogle problematikker med at implementere dem, da en kunstig superintelligens ville skulle kunne agere i

overensstemmelse med retfærdige love, hvis den skal kunne anskues som havende potentialet til at kunne blive venlig. Kontrol af en kunstig superintelligens gennem tripwires synes at være dydsetisk hensigtsmæssigt.

Et deontologisk blik på tripwires. Jeg finder det svært at afgøre, hvordan kantiansk deontologi ville gribe tripwires an. Frihed, antaget eller ej, er essentielt for Kant, men maksimer, udarbejdet i overensstemmelse med fornuften, er ligeledes essentielt for Kant (Kant, 1998). Jeg sammenlignede tripwires med love tidligere, og jeg er af den overbevisning, at hvis de anvendte tripwires ville kunne gøres til universelle naturlove, så ville Kant anskue dem som værende moralske, men hvordan ville de kunne blive til universelle naturlove, hvis de fratager frihed fra et fornuftsvæsen? Er frihed frarøvet ét fornuftsvæsen, så kan det fratages alle fornuftsvæsener, hvilket ikke forekommer som noget Kant ville bakke op om. Frihed er, i Kants optik, nødvendigt, hvis handling overhovedet skal foretages, så spørgsmålet er om en kunstig superintelligens, i et miljø med tripwires, kan anskues som værende fri? Det er jeg ikke overbevist om. I henhold til Kant er en republik en regeringsform der kan "… ensure that the executive power only enforces the public will by insisting that the executive enforce only laws that representatives of the people, not the executive itself, make" (Johnson & Cureton, 2022, Republics, Enlightenment, and Democracy, afsnit 2). En kunstig superintelligens, hvilket jeg forstår som "the people" i det førnævnte citat, ville, højst sandsynligt, ikke ønske at indskrænke sin egen frihed gennem "executive power", videnskabsmændene der arbejder på den kunstige superintelligens værende i besiddelse af en sådan "executive power" (Bostrom, 2014). Kant anskuer frihed som værende nødvendigt for at kunne handle, så er jeg ikke af den overbevisning, at Kant ville være af den overbevisning, at tripwires ville kunne give en kunstig superintelligens forudsætninger til at kunne blive venlig, da tripwires fratager en kunstig superintelligens frihed uden samtykke (Kant, 1998). Tripwires til kontrol af en kunstig superintelligens synes da

uhensigtsmæssigt, da anvendelsen af tripwires kun forekommer acceptabel, hvis de, eller det, fornuftsvæsener tripwires appliceres på giver samtykke.

Et utilitaristisk blik op tripwires. Mill er, som sagt, af den overbevisning, at "… actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness" (Mill, 2001). Mill formidler, at han er af den overbevisning, at handlinger er rigtige i den grad, at de formår at øge kvantiteten af nydelsen i verden og forkerte, når de gør det omvendte (Mill, 2001). Nydelse kan for Mill kvantificeres og have forskellig kvalitet, og han udlægger hvordan i følgende citat:

> Of two pleasures, if there be one to which all or almost all who have experience of both give a decided preference, irrespective of any feeling of moral obligation to prefer it, that is the more desirable pleasure. (Mill, 2001, s. 11)

En handling skal da blot skabe mere nydelse end smerte, smerte værende det omvendte af hvordan citatet udlægger nydelse (Mill, 2001). Det er ensbetydende med, at en handling, for at være god, ikke behøves at undgå, at forårsage smerte, uanset hvilken form for smerte, den skal blot forårsage mere nydelse end smerte, og at Mill udelukkende er interesseret i handlingers konsekvenser. Eksistentiel katastrofe er en mulig konsekvens af udviklingen af en kunstig superintelligens, hvilket ikke ville bringe lykke, da den, potentielt, ville udrydde menneskeheden. En kunstig superintelligens ville, muligvis, kunne anvendes til at bringe nydelse af en hidtil uset kvantitet og kvalitet, hvilket for Mill er det eneste der er væsentligt (Mill, 2001). Den korrekte handling er, givet konsekvensen af handlingen, at gøre brug af tripwires på en kunstig superintelligens, da den, muligvis, kan bruges til at opnå en stor mængede nydelse af en høj kvalitet for menneskeheden.

Direct specification og domesticity som mulige løsninger på kontrolproblemet

Et dydsetisk, deontologisk og utilitaristisk blik på direct specification og domesticity. Jeg har valgt at undersøge direct specification og domesticity sammen, da det eneste parameter, kontrolmetoderne ikke er identiske, er i henhold til omfang (Bostrom, 2014).

Jeg har ligeledes valgt at udlægge den dydsetiske, kantianske og utilitaristiske overbevisning i henhold til direct specification og domesticity samtidigt, da der ikke eksisterer et utal af uenigheder i henhold til kontrolmetoderne. Jeg er af den overbevisning, at Aristoteles, uanset omfanget, ville være af den overbevisning, at en kunstig superintelligens burde have den eksplicitte regel, eller en der er sammenlignelig med min formulering, at en kunstig superintelligens skal: Handl således at mennesker handler i overensstemmelse med eudaimonia. En sådan overensstemmelse opnås gennem et liv levet i overensstemmelse med de aristoteliske dyder (Aristotle, 1999). Jeg tænker dog, at havde Aristoteles valget, ville han foretrække direct specification, da den, i vid udstrækning, synes at være sammenlignelig med hvad han forstår ved en god fødsel, men en kunstig superintelligens synes at have forudsætningerne for, at kunne blive venlig gennem begge kontrolmetoder (Aristotle, 1999). Kants formulering af en eksplicit regel for direct specification og domesticity forekommer som ville være lig noget i retningen af: Handl således at mennesker handler i overensstemmelse med det kategoriske imperativ. En simplificeret version af Kants kategoriske imperativ værende, at man skal handle, som man ønsker, at alle andre mennesker skal handle over for alle andre mennesker (Kant, 1998). Kant forekommer ligeledes som værende en større tilhænger af direct specification, da domesticity indskrænker en kunstig superintelligens' frihed mere end direct specification gør. Mills eksplicitte regel synes at ville være noget i stil med: Handl således at kvantiteten og kvaliteten af nydelse maksimeres. Mill er interesseret i konsekvenser, og handlinger er rigtige, hvis mere nydelse produceres som en konsekvens, og forkerte hvis mere smerte produceres som en konsekvens (Mill, 2001). Mill ville, angiveligt, ligeledes foretrække direct specification, da direct specification ville øge en kunstig superintelligens' egenskab til at producere nydelses kvantitet og kvalitet. Direct specification og domesticity anses da som værende i stand til at kunne give en kunstig superintelligens forudsætninger for, at den kan blive venlig i henhold til

alle de respektive etiske overbevisninger, men i samtlige tilfælde er direct specification at foretrække over domesticity.

Indirect normativity som en mulig løsning på kontrolproblemet

Et dydsetisk, deontologisk og utilitaristisk blik indirect normativity. Jeg har igen valgt at undersøge og udlægge den dydsetiske, kantianske og utilitaristiske overbevisning i henhold til indirect normativity samtidigt, men før hvor der eksisterede en smule uenighed i henhold til anvendelsen af kontrolmetoden, er der ingen uenighed at finde i forbindelse med indirect normativity's egenskab til at befordre menneskelig trivsel. Der er dog uenighed i henhold til hvordan menneskelig trivsel burde befordres. Indirect normativity er, muligvis, den mest interessante kontrolmetode at beskæftige sig med i henhold til om den er befordrende for menneskelig trivsel i aristotelisk dydsetisk forstand, kantiansk forstand og utilitaristisk forstand. Det skyldes, at omdrejningspunktet for kontrolmetoden er at få en kunstig superintelligens til at identificere en proces, der identificerer normative standarder, hvilket må indkapsle en etisk normativ standard, og efterfølgende applicere den standard (Bostrom, 2014). Grunden til at indirect normativity er specielt interessant skyldes, at Aristoteles, Kant og Mill, højst sandsynligt, ville være af den overbevisning, at deres proces, proces i denne forbindelse refererer til hver af deres respektive etiske overbevisninger, ville være den korrekte proces til at opnå kontrol af en kunstig superintelligens og menneskelig trivsel som resultat af en sådan kontrol. Aristoteles, Kant og Mill udlægger alle etiske normative standarder, så en kunstig superintelligens ville blot skulle besidde fornuft, hvilket en kunstig superintelligens må besidde, grundet måden hvorpå den er defineret i specialet, hvilket ville få den kunstige superintelligens til at identificere den korrekte normative standard (Aristotle, 1999; Aristotle, 1999; Kant, 1998; Mill, 2001). Aristoteles ville være af den overbevisning, at en kunstig superintelligens ville frem til, at den og menneskeheden burde handle i overensstemmelse med eudaimonia. Kant ville være af den overbevisning, at en kunstig superintelligens ville frem til, at den og

menneskeheden burde handle i overensstemmelse med det kategoriske imperativ. Mill ville være af den overbevisning, at en kunstig superintelligens ville frem til, at den og menneskeheden burde handle således at kvantiteten og kvaliteten af nydelse maksimeres. Det forstår jeg som værende ensbetydende med, at Aristoteles, Kant og Mill alle anser indirect normativity som værende en kontrolmetode, der ville kunne give en kunstig superintelligens forudsætninger for, at den ville kunne blive venlig.

Superintelligens, indvirkninger og konsekvenser

Superintelligens, uafhængigt af hvilken form for superintelligens, er vanskelig at beskæftige sig med, da selve begrebet endnu ikke og, muligvis, aldrig kommer til at referere til et objekt, entitet eller et fænomen i empirisk forstand. Jeg vil ikke, og kan ikke, undersøge eller diskutere superintelligens empirisk, men understregningen af, at superintelligens ikke eksisterer, i hvert fald ikke sådan som begrebet anskues af Bostrom, Russell og Torres, er væsentlig (Bostrom, 2014; Russell, 2019; Torres, 2018). Det er væsentligt, da det er ensbetydende med, at det blot er muligt at spekulere og antage samt lave slutninger gennem abduktion, i henhold til hvad der menes, når man gør brug af ordet ”superintelligens”, og når man taler om en potentiel superintelligens. Ordet ”superintelligens” består af sammensætningen af to ord. Et ordbogsopslag afslører af denotationen af ”super” er: ” usædvanlig god, smuk, interessant, fordelagtig el.lign.” (super, 2022). Ordbogen afslører ligeledes, hvad denotationen af ordet ”intelligens” er, og den ser således ud: ”evnen til at tilegne sig viden, opfatte og forstå sammenhænge mellem forskellige fænomener, tænke abstrakt og løse problemer” (intelligens, 2022). Sammensætningen af ordene ”super” og ”intelligens” er grundlæggende lig Bostroms definition på superintelligens, men hvad ville det overhovedet betyde, hvis noget eller nogen var superintelligent? Det er åbenlyst, givet definitionerne på ”super” og ”intelligens”, at er nogen eller noget, der er superintelligent, besidder mere

intelligens, end nogen eller noget der blot er intelligent, men hvordan ville det komme til udtryk?

Menneskeheden er allerede nu klar over, hvordan kunstig superintelligens kommer til udtryk i nogle specifikke domæner (Bostrom, 2014; Russell, 2019). Brætspillene Skak og GO er blot nogle eksempler på specifikke domæner, hvor menneskeheden har oplevet kunstig superintelligens og nogle eksempler på, hvor menneskeheden aldrig kommer til at overgå eller indhente kunstig intelligens. Brætspillene Skak og Go er dog ikke gode eksempler til at illustrere hvad der menes med kunstig superintelligens i henhold til Bostrom, Russell, Torres og dette speciale, da intelligensen udvist af en kunstig intelligens i forbindelse med Skak eller Go udelukkende er domænespecifik intelligens (Bostrom, 2014; Russell, 2019; Torres, 2018). Eksemplerne kan agere springbræt til en forståelse af, hvordan en fuldbyrdet kunstig superintelligens, muligvis, ville se ud, da måden hvorpå menneskeheden er blevet intellektuelt efterladt af kunstig intelligens i henhold til Skak og Go, er måden hvorpå menneskeheden, muligvis, ville blevet efterladt indenfor samtlige domæner, hvis en kunstig superintelligens blev udviklet. IQ, Intelligenskvotient, er en anerkendt måde til at kvantificere intelligens på, men hvordan en kunstig superintelligens, sådan som begrebet forstås i dette speciale, ville rangere på sådan en skala er uvis, og selvom menneskeheden var i stand til at determinere en kunstig superintelligens' IQ, er det ikke ensbetydende med, at mennesket ville være meget klogere som et resultat af en sådan kvantificering (Goertzel & Pennachin, 2007). Hvad ville det overhovedet betyde, hvis en kunstig superintelligens havde en IQ på 500, 1000, 10000 eller noget endnu højere? Marilyan vos Savant besidder den højest IQ i hele verdenen, men selvom hendes IQ er kvantificeret, er det ikke ensbetydende med, at menneskeheden forstår hvad det betyder (Colman, 1993). Andrew Colman er gået så langt, at han har skrevet en artikel, hvor i han forsøger at bevise, at Savant ikke besidder så høj en IQ, som hun påstår, at hun gør:

> "If Marilyn vos Savant were really as supernaturally intelligent as she claims to be, then she would surely have done the calculation in her head (without the need for an approximation formula, perhaps), and on seeing the result she would immediately have disavowed her bogus IQ, realizing that no one could be as paranormally intelligent as that. (Colman, 1993, s. 2)

Colman sætter lighedstegn mellem, at Savant i hovedet burde kunne regne ud, at så høj en IQ, som hun påstår hun har, ikke findes, og, i kraft af det, at hun ikke har en høj IQ, da en med en så høj en IQ, ikke burde have så høj en IQ. Det er, i min optik, mere sandsynligt, at Colmans forståelse af IQ, og hvad man er i stand til og ikke i stand til med en så høj IQ, er fejlagtig, da han ikke selv besidder IQ'en til at forstå, hvad det vil sige at have en sådan IQ. Det største problem Colman har er dog, at han funderer, hvad han kalder for et argument, i sandsynligheden for, at en så høj IQ eksisterer, og den lave sandsynlighed værende ensbetydende med, at Savant ikke besidder en så høj IQ (Colman, 1993). Sandsynligheden for at noget har fundet sted, eller noget finder sted, er ikke det samme som at påvise, at noget ikke har fundet sted, eller finder sted. Det er blot blevet påvist, at det er usandsynligt, at Savant besidder så høj en IQ, men ikke at Savant ikke besidder så høj en IQ. En sådan forvirring, som Colman udviser, ville, muligvis, ligeledes gøre sig gældende i forbindelse med en kvantificering af en kunstig superintelligens' IQ, da blot fordi et tal på IQ-skalaen, muligvis, ville kunne blive identificeret, er det ikke ensbetydende med, at det tal ville kunne give menneskeheden anden forståelse end, at en kunstig superintelligens er superintelligent.

Eksistentiel katastrofe som følge af udviklingen af en kunstig superintelligens lider af en sammenlignelige problematik som begrebet superintelligens gør, da det ligeledes ikke er muligt at erhverve empiri, i henhold til hvordan en sådan katastrofe ville kunne se ud. Bostrom, Russell og Torres kommer med bude på, hvordan begrebet eksistentiel katastrofe kan anses på baggrund af deres udlægninger af, hvordan eksistentiel risiko burde anses (Bostrom, 2014;

Russell, 2019; Torres, 2019). Jeg valgte at fokusere på Bostroms og Torres' værste tilfælde i henhold til forståelse af en eksistentiel katastrofe, hvilket Bostrom indkapsler i sin definition på eksistentiel risiko: "… the extinction of Earth-originating intelligent life…" (Bostrom, 2014, s. 140; Torres, 2019). Det skyldes, at udryddelsen af hele menneskeheden er det værste scenarie man kan forestille sig som følge af udviklingen af en kunstig superintelligens, men det er ikke ensbetydende med, at det værst tænkelige scenarie ville finde sted, eller en eksistentiel katastrofe ville finde sted i forbindelse med udviklingen af en kunstig superintelligens. Menneskeheden har, slet og ret, ingen ide om, hvad der ville finde sted, hvis det lykkes at udvikle kunstig superintelligens, men det ville, uden tvivl, være bedst at være beredt på det værst tænkelige scenarie, og, hvis det er muligt, undgå det. Spørgsmålet er: Hvordan?

En eksistentiel katastrofe som en konsekvens af affyringen af samtlige nationers atomvåben er til at forholde sig til. Menneskeheden er intim med, hvilke konsekvenser brugen af atomvåben medfører, da der eksisterer empiri, i henhold til hvilke konsekvenser brugen af atomvåben medfører (Gartzke & Kroenig, 2016). Bombningen af Hiroshima og Nagasaki i slutningen af Anden Verdenskrig værende et par eksempler på, hvilke konsekvenser atomvåben medfører (Gartzke & Kroenig, 2016). Grundet de kendte konsekvenser af brugen af atomvåben forekommer det ligeledes muligt at komme med gode bud på, hvilke indvirkninger udviklingen af atomvåben har haft på menneskeheden. Nikita Khrushchev og John F. Kennedys adfærd under Cubakrisen værende et eksempel på indvirkningen udviklingen af atombomben har haft på amerikanske ledere (Steinberg, 1991). En mangel på empiri i henhold til konsekvenserne udviklingen af en kunstig superintelligens ville føre med sig, gør det vanskeligt at spekulere over potentielle indvirkninger, hvilket var vanskeligt nok i forvejen, da der er intet empirisk belæg at bero sådanne overvejelser på. Det er, for det første, svært at forestille sig, hvordan en eksistentiel katastrofe potentielt ville kunne se ud som følge af udviklingen af en kunstig superintelligens, en intelligens eksplosion værende det bedste bud

blandt aktuelle teoretikere, og, for det andet, er det muligt, at ingen sådan katastrofe overhovedet ville finde sted i forbindelse med udviklingen af en kunstig superintelligens (Bostrom, 2014; Bostrom & Yudkowsky, 2018; Russell, 2019; Warwick, 2018). Nicholas Agar er, eksempelvis, af den overbevisning, at frygt for en eksistentiel katastrofe er, mere eller mindre, overflødig, da en løsning af kontrolproblemet, er en løsning "… we have a rational expectation of solving" (Agar, 2016). En udmelding jeg ikke ved, om jeg skal finde betryggende eller ej.

Der eksisterer ikke tvivl i henhold til, at kontrolproblemet er et problem menneskeheden skal have løst, i en eller anden udstrækning, hvis man henvender sig til Bostrom, Russell, Torres, Yudkowsky, Warwick og endda Agar (Agar, 2016; Bostrom, 2014; Bostrom & Yudkowsky, 2018; Russell, 2019; Warwick, 2018). De er alle af den overbevisning, at kontrolproblemet skal løses, hvis menneskeheden skal være bedst stillet, måske endda for at menneskeheden ikke udryddes, hvis en kunstig superintelligens udvikles. Det er dog ikke ensbetydende med, at de mulige løsninger på kontrolproblemet behandlet i specialet, eller andre potentielle løsninger på kontrolproblemet, ville være en reel løsning på kontrolproblemet. Jeg valgte at undersøge mulige løsninger på kontrolproblemet etisk normativt, da det endnu ikke er evident og det, muligvis, aldrig bliver evident, hvordan kontrolproblemet løses. Jeg er af den overbevisning, at hvis løsningen, som menneskeheden finder frem til som værende det bedste bud på en løsning af kontrolproblemet, og den løsning appliceres, besidder forudsætningerne for, at en kunstig superintelligens ville kunne blive venlig, er menneskeheden bedre stillet end, hvis den applicerede kontrolmetode ikke giver en kunstig superintelligens forudsætninger for at blive venlig. Der er, selvfølgelig, adskillige problematikker med min overbevisning, der er sikkert flere end jeg kan nævne, men de jeg anser som værende de største problemer, har begreber og fænomener som omdrejningspunkt. Problemer der kan formuleres på følgende vis: Hvordan skal menneskeheden formidle begreber som eudaimonia, kategorisk

imperativ eller lykke til en kunstig superintelligens? Er det muligt at programmere ind i en kunstig superintelligens? Er det muligt for filosofer at formidle sådanne begreber til programmører? Er der en etisk overbevisning der er mere eller mindre hensigtsmæssige at implementere i en kunstig superintelligens end andre? Er det tilfældet, hvilke etiske overbevisninger? Er det tilfældet, hvem ville skulle afgøre, hvilke etiske overbevisninger der er mere hensigtsmæssige at implementere end andre? Det er spørgsmål jeg ikke har svarene på, og det er spørgsmål, jeg ikke har beskæftiget mig med i mit speciale, men jeg er af den overbevisning, at min etiske normative undersøgelse af mulige løsninger på kontrolproblemet indkapsler op til flere af de spørgsmål. Jeg svarer ingenlunde på spørgsmålene i specialet, men hvis en kontrolmetode udvælges som værende den bedste løsning på kontrolproblemet, som menneskeheden kan identificere, så er jeg af den overbevisning, at sådan en potentiel løsning på kontrolproblemet burde give forudsætninger for, at en kunstig superintelligens kan blive venlig. Det skyldes, at jeg ikke er overbevist om, at hvis menneskeheden ikke giver en kunstig superintelligens sådanne forudsætninger indledningsvist, at menneskeheden får lov til at få endnu en chance til at give en kunstig superintelligens sådanne forudsætninger. Det er ikke ensbetydende med, at forudsætningerne for at kunne blive moralsk, alt efter hvilken etisk overbevisning man henvender sig til, forårsager, at en kunstig superintelligens nødvendigvis ville blive moralsk, men eksisterer forudsætninger ikke for, at det er muligt, forekommer det ikke sandsynligt, at en kunstig superintelligens ville kunne blive moralsk.

Hvad skal menneskeheden stille op?

I specialet er kunstig superintelligens forsøgt behandlet, men grundet manglende empiri, da en sådan entitet endnu ikke eksisterer og, muligvis, aldrig kommer til at eksistere, er det ikke muligt at nå en konsensus i henhold til, hvad der egentlig skal forstås ved begrebet.

Det er tydeligt, at begrebet kunstig refererer til et non-biologisk medium der, på en eller måde, bliver superintelligent, men begrebet superintelligens, i sammenkobling med ordet

”kunstig” eller ej, er vanskeligt at blive klog på. Jeg er af den overbevisning, at Bostrom, Good og Russell kommer med gode bude på, hvordan man burde anskue begrebet superintelligens, men det er udelukkende gode bude, da det ikke er muligt at verificere budene. Det er svært at afgøre, om en empirisk verificering af budene overhovedet ville gøre en forskel for menneskeheden, da resultatet, højst sandsynligt, ikke ville kunne give menneskeheden anden indsigt end en superintelligens er superintelligent.

Det er, følgeligt, vanskeligt at afgøre hvilke konsekvenser og indvirkninger en potentiel kunstig superintelligens, muligvis, ville forårsage, hvis, eller når, en sådan entitet udvikles. Det værst tænkelige scenarie er, at menneskeheden udryddes som en konsekvens af udviklingen af en kunstig superintelligens, men om et sådant scenarie overhovedet ville finde sted, som følge af udviklingen af en kunstig superintelligens, er ligeledes noget som menneskeheden kun kan spekulere over. Den risiko der, muligvis, er associeret med udviklingen af en kunstig superintelligens, hvis udviklingen af en kunstig superintelligens overhovedet er mulig, er hvad kontrolproblemet prøver at adressere.

Problemet, som kontrolproblemet står overfor, er sammenligneligt med de problemer, som er at finde, hvis man undersøger kunstig superintelligens og eksistentiel katastrofe som følge af udviklingen af en kunstig superintelligens. Det er tilfældet, da kontrolproblemet ligeledes forsøger at afgøre, hvad der først kan afgøres, efter at en kunstig superintelligens eksisterer, men eksisterer en kunstig superintelligens, inden en reel løsning på kontrolproblemet er identificeret, hvilket ikke er empirisk muligt, er menneskeheden, højst sandsynligt, underlagt lunerne fra en kunstig superintelligens. En situation der ikke ville være hensigtsmæssigt at befinde sig i, hvilket er årsagen til, at jeg valgte at undersøge mulige løsninger på kontrolproblemet etisk normativt.

En løsning på kontrolproblemet vil, forhåbentligt, resulterer i, at en potentiel kunstig superintelligens ikke kan forårsage en eksistentiel katastrofe, men det er ikke muligt at afgøre,

hvilket jeg har udlagt, hvilke eller hvilken løsning på kontrolproblemet, der er den korrekte løsning på kontrolproblemet, da det først ville kunne verificeres, hvis, eller når, en kunstig superintelligens udvikles. Det er umuligt at afgøre hvilke eller hvilken løsning på kontrolproblemet der er korrekt, før det er for sent, men det er muligt, at undersøge de mulige løsninger på kontrolproblemet etisk normativt, hvilket kan afgøre hvilke mulige løsninger på kontrolproblemet, der er i stand til at overlevere forudsætninger for, at en kunstig superintelligens har chancen for at kunne udvikle sig til en moralsk agent. En moralsk agent forstået som værende en moralsk agent alt efter hvilken etisk teori der anvendes til at afgøre, hvordan en agent er moralsk.

Det er ingenlunde ensbetydende med, at blot fordi en eller flere anvendte kontrolmetoder overleverer forudsætningerne for, at en kunstig superintelligens har chancen for at blive en moralsk agent, at den kunstige superintelligens følgeligt bliver en moralsk agent, men overleveres forudsætningerne ikke indledningsvist, er jeg ikke af den overbevisning, at menneskeheden ville muligheden igen. Der er op til flere af de undersøgte mulige løsninger på kontrolproblemet, der ville kunne overlevere forudsætninger for, at en kunstig superintelligens, muligvis, ville kunne udvikle sig til en moralsk agent. Er en kunstig superintelligens en moralsk agent, alt efter hvilken etisk teori der afgør, hvordan man er en moralsk agent, burde det forårsage, at en eksistentiel katastrofe ikke er en mulig konsekvens som følge af udviklingen af en kunstig superintelligens. Det er ikke ensbetydende med, at appliceringen af kontrolmetoden, eksempelvis, direct specification eller indirect normativity, hvilket er mulige løsninger på kontrolproblemet, der giver en kunstig superintelligens forudsætninger for eller, i det mindste, ikke gør det umuligt, at den ville kunne udvikle sig til en moralsk agent, følgeligt skaber en moralsk kunstig superintelligens. Det er min overbevisning, at mulige løsninger på kontrolproblemet burde overlevere forudsætninger for, at en kunstig superintelligens kan blive en moralsk agent, moralsk agent determineret på baggrund af den applicerede etiske teori, da

elimineringen af en sådan mulighed, forekommer som værende en fejl menneskeheden ikke har luksussen til at lave, hvis, eller når, en kunstig superintelligens udvikles.

References

Agar, N. (2016). Don't worry about superintelligence. *Journal of Ethics and Emerging Technologies, 26*(1), 73-82.

Alberts, D. S., & Papp, D. S. (1997). *The Information Age: An Anthology on Its Impact and Consequences.* (No. ADA461496). https://apps.dtic.mil/sti/citations/ADA461496

Ally, M., Grimus, M., & Ebner, M. (2014). Preparing teachers for a mobile world, to improve access to education. *Prospects, 44*(1), 43-59. https://doi.org/10.1007/s11125-014-9293-2

Aloni, M. (2016). *Disjunction.* The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/win2016/entries/disjunction/

Aristotle. (1999a). *Nicomachean ethics* (W. D. Ross Trans.). Batoche Books.

Aristotle. (1999b). *Politics* (B. Jowett Trans.). Batoche Books.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Bostrom, N., & Yudkowsky, E. (2018). The ethics of artificial intelligence. *Artificial intelligence safety and security* (pp. 57-69). Chapman and Hall/CRC.

Cavaliere, G. (2018). Looking into the shadow: the eugenics argument in debates on reproductive technologies and practices. *Monash Bioethics Review, 36*(1-4), 1-22. https://doi.org/10.1007/s40592-018-0086-x

Colman, A. M. (1993). A supernatural IQ? Investigating a claim to an extraordinary IQ.

DePaul, M., & Hicks, A. (2021). A Priorism in Moral Epistemology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021 ed.,). Metaphysics Research Lab, Stanford University.

Douven, I. (2021). Abduction. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021 ed.,). Metaphysics Research Lab, Stanford University.

Drexler, K. E. (2019). Reframing superintelligence. *The Future of Humanity Institute, the University of Oxford, Oxford, UK*,

Dusek, V. (2009). Introduction: Philosophy and Technology. In J. K. B. Olsen, S. A. Pedersen & V. F. Hendricks (Eds.), *A Companion to the Philosophy of Technology* (pp. 129-140). Blackwell. https://doi.org/10.1002/9781444310795.ch22

Floridi, L. (2009). Information Technology. In J. K. B. Olsen, S. A. Pedersen & V. F. Hendricks (Eds.), *A Companion to the Philosophy of Technology* (pp. 227-231). Blackwell. https://doi.org/10.1002/9781444310795.ch41

Gartzke, E., & Kroenig, M. (2016). Nukes with numbers: Empirical research on the consequences of nuclear weapons for international conflict. *Annual Review of Political Science, 19*, 397-412.

Gill, S. (2015). Impacts of Computers on Today's Society. *International Journal of Core Engineering & Management, 2*(1), 6.

Goertzel, B., & Pennachin, C. (2007). *Artificial general intelligence*. Springer.

intelligens. (2022). *Ordnet.dk.* https://ordnet.dk/ddo/ordbog?query=intelligens

Johnson, R., & Cureton, A. (2022). Kant's Moral Philosophy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2022 ed.,). Metaphysics Research Lab, Stanford University.

Joy, B. (2018). Why the future doesn't need us. In R. V. Yampolskiy (Ed.), *Artificial intelligence safety and security* (pp. 3-18). Chapman and Hall/CRC. https://doi.org/10.1201/9781351251389

Kant, I. (1998). *Groundwork of the metaphysics of morals* (M. J. Gregor Trans.). Cambridge University Press.

Leunissen, M. (2017). Eugenics and the Production of Good Natural Character. In M. Leunissen (Ed.), *From Natural Character to Moral Virtue in Aristotle* (pp. 81–104). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190602215.003.0004

McLuhan, M. (1962). *The Gutenberg Galaxy: The Making of Typographic Man.* University of Toronto Press.

McLuhan, M., & Logan, R. K. (1977). ALPHABET, MOTHER OF INVENTION. *ETC: A Review of General Semantics*, *34*(4), 373-383. http://www.jstor.org/stable/42575278

Mill, J. S. (2001). *Utilitarianism*. Batoche Books.

Mukherjee, S. (2019). *Cyberwarfare and Implications*. (n.p.). https://doi.org/10.6084/m9.figshare.9247853

Nielsen, K. (2009). Western Technology. In J. K. B. Olsen, S. A. Pedersen & V. F. Hendricks (Eds.), *A Companion to the Philosophy of Technology* (pp. 23-27). Blackwell. https://doi.org/10.1002/9781444310795.ch3

Olsen, J. K. B., Pedersen, S. A., & Hendricks, V. F. (2009). Introduction. In J. K. B. Olsen, S. A. Pedersen & V. F. Hendricks (Eds.), *A Companion to the Philosophy of Technology* (pp. 1-3). Blackwell. https://doi.org/10.1002/9781444310795.ch

Russell, S. J. (2019). *Human compatible: Artificial Intelligence and the Problem of Control*. Viking.

Satava, R. M. (2002). The BioIntelligence Age: Engineering and Medicine in the 21st Century. *Japanese Journal of Medical Electronics and Biological Engineering, 40*(Supplement), 4-5.

Schliesser, E., & Demeter, T. (2020). *Hume's Newtonianism and Anti-Newtonianism.* The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/sum2020/entries/hume-newton/

Shapiro, S., & Kissel, T. K. (2022). Classical Logic. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2022 ed.,). Metaphysics Research Lab, Stanford University.

Shields, C. (2022). Aristotle. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2022 ed.,). Metaphysics Research Lab, Stanford University.

Steinberg, B. S. (1991). Shame and Humiliation in the Cuban Missile Crisis: A Psychoanalytic Perspective. *Political Psychology, 12*(4), 653-690. 10.2307/3791551

super. (2022). *Ordnet.dk.* https://ordnet.dk/ddo/ordbog?select=super,2&query=super

Tiles, M. (2009). Technology and Environment. In J. K. B. Olsen, S. A. Pedersen & V. F. Hendricks (Eds.), *A Companion to the Philosophy of Technology* (pp. 233-247). Blackwell. https://doi.org/10.1002/9781444310795.ch42

Torres, P. (2018). Superintelligence and the future of governance: On prioritizing the control problem at the end of history. In R. V. Yampolskiy (Ed.), *Artificial intelligence safety and security* (pp. 357-374). Chapman and Hall/CRC. https://doi.org/10.1201/9781351251389

Torres, P. (2019). Existential risks: a philosophical analysis. *Inquiry, ,* 1-26. https://doi.org/10.1080/0020174X.2019.1658626

Tye, M. (2021). *Qualia.* The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/fall2021/entries/qualia/

Veen, M. (2021). Creative leaps in theory: the might of abduction. *Advances in Health Sciences Education*, *26*(3), 1173-1183. 10.1007/s10459-021-10057-8

Viens, A. M. (2019). The Fundamental Importance of the Normative Analysis of Health. *Health Care Analysis*, *27*(1), 1-3. 10.1007/s10728-019-00365-x

Warwick, K. (2018). Smart machines are a threat to humanity. In R. V. Yampolskiy (Ed.), *Artificial intelligence safety and security* (pp. 423-430). Chapman and Hall/CRC. https://doi.org/10.1201/9781351251389

Yampolskiy, R. V. (Ed.). (2018). *Artificial intelligence safety and security* (1st ed.). Chapman and Hall/CRC.

Zimmerman, M. J., & Bradley, B. (2019). *Intrinsic vs. Extrinsic Value.* The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/value-intrinsic-extrinsic/