JÖNKÖPING UNIVERSITY
School of Education and Communication

”Jag tänker, alltså är jag en maskin”

En kvalitativ undersökning av hur AI omfattas av moraliskt ansvar

KURS: Examensarbete med opposition 15 hp
PROGRAM: Ämneslärarprogrammet
FÖRFATTARE: Erik Hallberg
EXAMINATOR: Peter Carlsson
TERMIN: VT19

Sammanfattning

”Jag tänker, alltså är jag en maskin” – en kvalitativ undersökning av hur AI omfattas av moraliskt ansvar

Förnamn Efternamn: Erik Hallberg

Antal sidor: 42

Uppsatsen är en komparativ kvalitativ analys av artiklar, böcker och föreläsningar som avser moraliskt ansvar för artificiell intelligens. Den teoretiska ansatsen tar intryck av moralfilosofen Peter Singers definition av artdiskriminering och moraliskt ansvar enligt religionsvetaren och filosofen Hans Jonas. Syftet med följande uppsats är att undersöka hur artificiell intelligens omfattas av moraliskt ansvar.
Källorna utgörs av moralfilosofer, fysiker och futurologer.
Frågeställningarna lyder som följer:

- Hur omfattas artificiell intelligens av vårt moraliska ansvar enligt teoretiker som resonerar om ämnet?
- Vilka likheter respektive skillnader finns mellan de olika författarna?
- Vad kan dessa likheter respektive skillnader antas bero på?
- Hur kan slutsatserna liknas vid Peter Singers teori om artdiskriminering?

Vid analys av materialet jämförs författarnas hållning komparativt och kvalitativt vad gäller moraliskt ansvar för att sedan jämföras med Singers definition av artdiskriminering. Det framgår att flertalet författare skildrar moraliskt ansvar utifrån mänsklighetens välbefinnande, att ansvaret ligger i ansvarsfull utveckling och eventuella dilemman. Ett fåtal författare skriver om moraliskt ansvar genom rättigheter för enskilda individer och vissa använder sig av normativa etiska teorier för att styrka sin position.

Sökord: AI, moraliskt ansvar, artdiskriminering

Postadress	**Gatuadress**	**Telefon**	**Fax**
Högskolan för lärande och kommunikation (HLK) Box 1026	Gjuterigatan 5	036-101000	036-162585

Innehållsförteckning

1. Inledning

Den tekniska utvecklingen de senaste decennierna har förändrat människors levnadssätt i grunden på flera plan. Det verkar också som att den kommer fortsätta i en överskådlig framtid där nu även artificiell intelligens, AI även kallat, kommer att ta en allt större plats i både det offentliga och privata livet. Hur detta ska hanteras rent praktiskt väcker flertalet frågor som hur arbetsmarknaden frågar och om människan genom sådan teknik på sikt kan komma att ersätta sig själv. Det återfinns även intresse i att undersöka moraliska aspekter av den nya AI-teknologin. Flertalet normativa etiska teorier berör även individer som nödvändigtvis inte är mänskliga, mot bakgrund av detta finns intresse i att undersöka hur AI-robotar behandlas enligt dessa teorier och jämföra dessa. Vad gäller utveckling av AI väcks genast frågor om hur och när AI kan bli medvetna individer och huruvida sådana skulle interagera med människor. Dessutom kan själva utvecklingen av AI utgöra grunden för moralisk diskussion, exempelvis om det är moraliskt att utveckla artificiellt intelligenta individer över huvud taget. Detta kan sammanfattas i moraliskt ansvar som vi människor har inför den teknologiska utvecklingen av AI och huruvida enskilda artificiellt intelligenta individer kan omfattas av moraliskt ansvar. Det uppstår även frågor kring om utvecklad AI har moraliskt ansvar för oss människor. Att undersöka moraliskt ansvar denna framtida aktör anses nödvändigt då utvecklingen kommer leda till framtida handlingar och utvecklingen självt kan anses vara en handling vilka kommer vara i behov av ordning och värdering vilket är vad etik och moral innebär.[1]

Det finns flertalet moralfilosofer och etiker som resonerar kring rättigheter och moralisk hänsyn mellan människor och icke-mänskliga varelser. Moralfilosofiska resonemang om artificiell intelligens återfinns då handlingar med etiska konsekvenser värderas, mot bakgrund av detta finns ett intresse att komparativt jämföra huruvida etiska principer används godtyckligt eller konsekvent genom att testa dem mellan andra individer än människor där de flesta anser att moraliskt ansvar bör ges. Människans samverkan med andra livsformer har tidigare beskrivits av Tom Regan och Peter Singer som skildrar samverkan mellan människor och djur som artdiskriminering.[2] Människans moraliska

[1] Jonas, Hans. *Ansvarets etik–utkast till en etik för den teknologiska civilisationen*. Göteborg: Bokförlaget Daidalos AB. 1991, 56.

[2] Regan, Tom. *Djurens rättigheter–en filosofisk argumentation*. Nora: Nya doxa. 1999, 45–46.

ansvar för djur har pågått under tusentals år medan AI är ett nytt fenomen, detta gör att det finns intresse i att undersöka även moraliskt ansvar för AI. Eftersom artificiellt intelligenta individer är att betrakta som en egen art kan detta användas för att undersöka huruvida människan väljer att hantera moraliskt ansvar även gentemot AI. Detta kan bidra med ny kunskap om människan självt genom att studera vårt ansvar gentemot andra.[3]

2. Syfte och frågeställningar

Syfte

Uppsatsens syfte är att undersöka hur den samtida moralfilosofin skildrar människans ansvar för artificiell intelligens och jämföra med Peter Singers teori om antidiskriminering.

Frågeställningar

Utifrån tidigare nämnda syfte formuleras de frågeställningar som följer:

- Hur omfattas artificiell intelligens av vårt moraliska ansvar enligt teoretiker som resonerar om ämnet?
- Vilka likheter respektive skillnader finns mellan de olika författarna?
- Vad kan dessa likheter respektive skillnader antas bero på?
- Hur kan slutsatserna liknas vid Peter Singers teori om antidiskriminering?

3. Material

Under rubriken materialbeskrivningen presenteras en beskrivning av materialet följt av en materialdiskussion där materialet problematiseras och diskuteras.

Materialbeskrivning

Uppsatsens material består i böcker skrivna av filosofer och teoretiker som företräder olika uppfattningar mellan människor och AI.

Virginia Dignum är professor vid institutionen för datavetenskap och arbetar vid Umeå universitet. Hon skrev 2018 artikeln ”Ethics in artificial intelligence: introduction to the special issue” som behandlar tre artiklar som berör ämnet och sammanställer det som hon anser är den rådande positionen gällande moraliskt ansvar mellan människor och artificiell

[3] Singer, Peter. *Djurens frigörelse*. Nora: Bokförlaget Nya doxa. 1999, 39.

intelligens. Dignum resonerar för hur ansvaret mellan människor och AI-robotar fördelas när synen på teknologisk utveckling skiftar från att betrakta tekniska instrument som autonomiskt agerande agenter. Dignum menar att denna nya utveckling kommer behöver diskuteras etiskt för att nå ett system som kan bygga tillit och respekt för mänskliga rättigheter och på så sätt undvika dystopiska framtidsvisioner och istället harmonisera relationen mellan människa och maskin.[4]

Stefan Larsson, Mikael Anneroth, Anna Felländer, Li Felländer-Tsai, Fredrik Heintz och Rebecka Cedering Ångström som forskar inom datadrivna marknader, AI, digital juridik, maskininlärning, social förändring, datavetenskap och sociala perspektiv på IKT skrev i mars 2019 rapporten: ”Hållbar AI: inventering av kunskapsläget för etiska, sociala och rättsliga utmaningar med artificiell intelligens” där de sammanställde rådande utmaningar gällande artificiell intelligens och dess interaktion med mänskligheten. De utmaningar som tas upp är etiska, sociala och rättsliga utmaningar med artificiell intelligens där de skildrar rapporten som en inventering av rådande kunskapsläge. Rapporten använder sig av en bibliometrisk analys för att fördjupa ovan nämnda problemområden gällande AI.[5]

Fysikern Max Tegmark skrev 2017 boken *Liv 3.0: Att vara människa i den artificiella intelligensens tid* där han resonerar kring hur artificiell intelligens kommer förändra samhället och utmaningar med detta. Tegmark menar att AI inte längre är en science fiction-utopi utan att det återfinns i en överskådlig framtid. Han resonerar kring människans förhållande till AI och hur världen kan struktureras upp på ett godtagbart sätt.[6]

Dorna Behdadi som är doktorand i praktisk filosofi vid Göteborgs universitet höll i början av april 2019 en föreläsning gällande etik och medvetenhet som heter *Conscious AI: A moral dilemma* via TEDx talks där hon resonerar kring medvetenhet och moraliskt ansvar mellan AI och människor. Behdadi problematiserar människans anspråk gällande teknologisk utveckling och medvetenhet mellan individer samt hur detta kan avgöras. Behdadi nämner bland annat att medvetenhet enbart går att förutsätta för den enskilda individen medan det bygger på antaganden gällande andra än sig själv.[7]

[4] Dignum, Virginia. *Ethics in artificial intelligence: introduction to the special issue.* Ehics and information technology. 2018.
[5] Larsson, Stefan. Anneroth, Mikael. Felländer, Anna. Felländer-Tsai, Li. Heintz, Fredrik. Cedering, Rebecka. Ångström. *Hållbar AI: inventering av kunskapsläget för etiska, sociala och rättsliga utmaningar med artificiell intelligens.* Stockholm: AI sustainability center. 2019.
[6] Tegmark, Max. Liv 3.0: *Att vara människa i den artificiella intelligensens tid*. Stockholm: Volante. 2017.
[7] Behdadi, Dorna. *Conscious AI: A moral dilemma*. TedxGöteborg. 2019.

Filosofiprofessorn vid Canterbury University Jack Copeland skrev 1993 boken *Artificial Intelligence: A philosophical introduction*. Copeland resonerar kring hur utvecklingen av AI pågått sedan 1950-talet och hur den tekniska utvecklingen och filosofi relaterar till denna. Copeland ifrågasätter medvetenhet och fri vilja bland maskiner där han sedan för en vidare diskussion kring huruvida människor och dess hjärnor kan fungera på ett sätt som även maskiner gör. Copeland vänder således på resonemanget kring om robotar kan vara människor och problematiserar om människor kan vara robotar.[8]

Nick Bostrom är en filosof som 2016 skrev boken *Superintelligence: Paths, dangers and strategies.* Bostrom skildrar i boken hur analytiskt tänkande kan hjälpa mänskligheten att navigera i en framtid där maskiner är intelligenta, medvetna och aktiva parter i världen. Bostrom frågar även om mänsklighet och vilka utmaningar som vi möter i en framtid där artificiell superintelligens kommer samverka med människor.[9] Bostrom har även skrivit artikeln *Ethical issues in advanced Artificial intelligence* 2011 där han granskar och analyserar risker och möjligheter med utvecklandet av AI vad gäller moral och etik. Bostrom resonerar även kring etiskt relevanta skillnader mellan människor och maskiner och vad som anses relevant för moraliskt ansvar, både för AI och gentemot.[10]

Professorn Mathias Risse skrev 2018 en artikel, ”Human rights and Artificial intelligense – An urgently needed agenda”, rörande AI och mänskliga rättigheter. Han har tidigare publicerat verk rörande politisk och social filosofi, Risse menar att utvecklingen av artificiellt medvetna individer utmanar tidigare föreställningar om mänskliga rättigheter. Han anser att vi behöver problematisera den hierarki av moraliskt ansvar som de flesta bär med sig där människan placeras högst och således erhåller flest rättigheter, denna problematisering anses nödvändig enligt Risse då AI potentiellt kan vara bärare av egenskaper vi människor anses avgörande för vår plats i hierarkin. Dessutom resonerar han kring att AI-robotar möjligtvis kan vara innehavare av dessa egenskaper i större utsträckning än oss. Risse redogör för och problematiserar kring vad mänskliga rättigheter innebär och resonerar kring vad som vara karaktärsdrag som berättigar moraliskt ansvar.[11]

[8] Copeland, Jack. *Artificial Intelligence: A philosophical introduction.* New Jersey*:* Blackwell Publishing. 1993.
[9] Bostrom, Nick. *Superintelligence: Paths, dangers and strategies*. Oxford: Oxford University Press. 2016.
[10] Bostron, Nick. *Ethical issues in advanced Artificial intelligence.* Science Fiction and Philosophy: From Time Travel to Superintelligence. Oxford: Oxford University. 2003.
[11] Risse, Mattias. *Human rights and Artificial intelligense – An urgently needed agenda.* Carr center for human rights policy. Cambridge, Massachusetts. 2018.

AI-forskaren Eleizer Yudkowsky skrev 2008 ett kapitel i Nick Bostroms verk *Global catastrophic risks* som heter *Artificial intelligence as a positive and negative factor in global risk* där han resonerar kring människans ansvar gällande utveckling av AI och han problematiserar dessutom kring intelligens och förhållandet mellan människor och robotar. Yudkowsky skildrar vidare vänlig AI där han resonerar kring hur människor och robotar kan samverka på ett sätt som inte riskerar att landa i skräckscenarion där vi riskerar att skapa lidande och destruktivitet. Han menar att resonemang kring ämnet är av stor prioritet då insatserna är långtgående.[12]

Bill Hibbard, som är vetenskapsman och forskar kring intelligenta maskiner vid University of Wisconsin–Madison skrev en bok 2014 som berör relationen mellan människa och AI. I boken *Ethical artificial intelligence* sammanfattar Hibbard sina tankar och teorier rörande artificiellt intelligenta individer och berör både teknologisk utveckling och etiska ställningstaganden. Han beskriver risker, möjligheter och förhållningssätt mellan människor och AI samtidigt som han förespråkar en utilitaristisk hållning vad gäller medvetna AI-individer.[13]

Paul Conrad Samuelsson skrev 2019 en artikel i ”Filosofisk tidskrift som heter Artificiella medvetanden är vår största etiska risk” där han resonerar kring mänskligheten och förhållandet till artificiellt medvetna individer. Samuelsson använder sig av en konsekvensetisk teori för att påtala de risker som återfinns med AI, både vad gäller människor och AI-individer själva. Han skiljer riskerna vad gäller människor och AI samtidigt som han på en metanivå resonerar kring människans tekniska framsteg och verktyg för att hantera etik i en tid av teknologisk utveckling.[14]

Micheal. R. LaChat, som är professor i teologisk etik, skrev en artikel i AI Magazine 1986 där han resonerade kring personlig AI och moral rörande detta. LaChat problematiserar i artikeln moraliska överväganden som dualism och rättigheter, möjligheten att skapa en AI-person och moraliskt ansvar gentemot och gällande AI-personer. LaChat skildrar moraliskt ansvar mot bakgrund av FN:s deklaration om mänskliga rättigheter och ifrågasätter

[12] Yudkowsky, Eleizer. *Artificial intelligence as a Positive and Negative factor in global risk.* I Global catastrophic risks. Bostrom, N. Crikovic, M. New York: Oxford university press. 2008.
[13] Hibbard, Bill. *Ethical artificial intelligence*. Berkeley: Space Science and Engineering Center
University of Wisconsin–Madison
and
Machine Intelligence Research Institute. 2014.
[14] Samuelsson, Paul–Conrad. *Artificiella medvetanden är vår största etiska risk*. Filosofisk tidskrift. Stockholm: Thales förlag. 2019.

huruvida AI-personer bör tjäna mänskligheten eller om detta kan betraktas som slaveri. Hans artikel problematiserar även mänskliga rättigheter på en metanivå där han utreder om det är det mänskliga i mänskliga rättigheter som ger rättigheter eller om det är andra variabler.[15]

Materialdiskussion

Valda teoretikers texter anses väl avvägda då de behandlar relationen mellan människor och artificiell intelligens på ett eller annat sätt samt att de resonerar kring etik rörande denna relation.

Materialet inbegriper även flertalet olika åsikter och utgångspunkter, de olika författarna har dessutom flera olika ingångar vad gäller förhållandet mellan människor och artificiell intelligens. De olika författarna har följaktligen väldigt olika ingångspunkter till ämnet vilket anses göra uppfattningen bred och omfattande snarare än begränsad då författaren inom enbart ett fält skulle innebära en uppsats som undersöker enbart vissa aspekter. En undersökning med författare från ett fält skulle vara lämpligt vid undersökningar av problem som är väl etablerade i rådande samhälle, detta anser jag inte att förhållandet mellan artificiell intelligens och människor är utan det är i behov av att problematiseras i samverkan med att det utvecklas. Detta gör att materialet lämpar sig väl för att möta undersökningens syfte samtidigt som det anses vara i behov av problematisering då det berör en relation som är komplex, under utveckling och något vars relevans successivt ökar i takt med den tekniska utvecklingen.

Det bör även sägas att materialet begränsas av ovan nämnda aktualitet då de som skriver om relationen mellan människor och artificiell intelligens främst är kopplade till den tekniska utvecklingen och/eller etik och moral. Detta begränsar urvalet eftersom artificiell intelligens ännu inte tagit den plats i samhället som den förmodas ha i framtiden. Som jämförelse kan det sägas att människans förhållande till varandra och även icke-mänskliga djur diskuterats vad gäller moraliskt ansvar under en längre period medan diskussionen om artificiell intelligens och människor är något nytt.

4. Teoretisk anknytning

[15] LaChat, Micheal. R. *Artifficial intelligence and ethics: An exercise in the moral imagination*. The AI magazine. Vol 7. Nr 2. 1986.

Vad gäller moraliskt ansvar så utgår uppsatsen från filosofen Hans Jonas *ansvarsteori* som avser ansvar som förutsätter kausal makt, en gärningsman är således ansvarig för sin handling och hålls skyldig botgörelse. Den skapade skadan bör således bestraffas och gottgöras samtidigt även om följden varit avsedd eller förutsedd, det som räcker för att den ska betraktas ansvarig är alltså att den ansvarige är en aktiv part och orsak till utfallet.[16] Jonas skiljer legalt och moraliskt ansvar, även om de anses härledas till varandra, genom att skildra ansvariga överträdelser som behov av gottgörelse medan legala övertramp ger skuld och den aktive betraktas som skyldig respektive inte skyldig. Jonas menar dock att båda gärningarna berör en ansvarig aktiv parts handlande gentemot andra. Ansvar avser utomstående, eller agenten självt, men det är inte ansvar i sig självt som ett ändamål utan är istället ett påbud vilket avser kausalt handlande bland individer. Jonas menar att ansvar på så sätt är en förutsättning för moral snarare än moral i sig självt.[17]

Något som gör Jonas definition av ansvar väl avvägd anses vara att den avser teknologisk utveckling och etik, något som uppsatsens syfte om moraliskt ansvar för artificiell intelligens tangerar. Jonas menar att teknik förändrar världen i och med att det inbegriper samhällets och naturens villkor och troligtvis kommer påverka framtida företeelser på ett exponentiellt sätt. Hans uppfattning är också att tanken om teknologi som framstegsbringande bidrar till förbättring av den allmänna moralen där den dessutom tar större plats.[18] Det är Jonas definition av moraliskt ansvar som undersöks i uppsatsen, detta yttrar sig genom att det är hans definition som teoretiker menar att människan har moraliskt ansvar för respektive inte har ansvar för.

Vidare tar den teoretiska ansatsen intryck av filosofen Peter Singers användning av begreppet *artdiskriminering* vilket avser att mänskliga varelser orsakar icke-mänskliga varelser stor mängd lidande trots att det inte behövs. Detta lidande orsakas dessutom trots att det strider mot principen om lika hänsyn till önskningar och att de flesta inte skulle acceptera att själva utsättas för lidande som inte är nödvändigt. Denna motsägelse och godtyckligt handlande med konsekvenserna som lidande, trots att det strider mot fundamental princip om moralisk hänsyn efter preferenser, är enligt Peter Singer förtryck enbart baserat på arttillhörighet. Detta artförtryck används dessutom godtyckligt där människor accepterar arttillhörighet som referens för moraliskt ansvar medan det inte

[16] Jonas, 121–152.
[17] Ibid., 153.
[18] Ibid., 257–259.

accepterats om förhållandena varit omvända. De flesta människor skulle nog inte acceptera att inte visas moralisk hänsyn vilket leder till stora mängder lidande med enbart hänvisning till arttillhörighet.[19] I uppsatsen yttrar sig detta genom att teoretikernas argument kommer granskas för att skönja om det återfinns drag av artdiskriminering i resonemang kring artificiell intelligens. Singers tydning av begreppet *artdiskriminering* lämpar sig väl då det avser relationen mellan mänskliga och icke-mänskliga varelser och det synliggör godtycklighet och logisk inkonsekvens.

5. Metod

Metodbeskrivning

Metoden är att betrakta som en kvalitativ komparativ analys av filosofiska resonemang rörande moraliskt ansvar mellan människor och artificiellt intelligenta individer. I uppsatsens resultatredovisning tolkas olika teoretikers förhållningssätt till moraliskt ansvar mellan människor och AI-individer på ett kvalitativt sätt för att visa på skillnader och likheter dem emellan. Den kvalitativa metoden yttrar sig genom att sträva efter att synliggöra, utöver likheter och skillnader teoretiker emellan, även efter tidigare icke-kända företeelser samt hur dessa yttrar sig i olika sammanhang som bland populationer eller situationer. Kvalitativa metoder leder till synliggörande av strukturer och/eller variationer mellan olika fenomen, resonemang och innebörder.[20] Vad gäller uppsatsen så används den kvalitativa metoden i resultatredovisningen genom att visa på hur olika teoretiker beskriver moraliskt ansvar mellan AI-individer och människor utifrån tidigare nämnda definition av moraliskt ansvar och analyser om det återfinns drag av artdiskriminering bland de resonemang som förs.

Metodens komparativa element märks i uppsatsen främst genom jämförandet mellan teoretikerna där likheter och skillnader synliggörs och jämförs. Den komparativa metoden används för att synliggöra likheter och skillnader mellan olika fenomen, men även för att beskriva, förklara och förutsäga.[21] Komparativa metoder utmärks av att de evaluerar egenskaper som finns mellan det som undersöks vilket sedan kan jämföras mellan

[19] Singer, 233.
[20] Starrin, Bengt. Svensson, Per–Gunnar. *Kvalitativ metod och vetenskapsteori*. Lund: Studentlitteratur AB.2009, 23.
[21] Denk, Thomas. *Komparativ metod–förståelse genom jämförelse*. Lund: Studentlitteratur AB. 2002, 28.

varandra. Det behöver minst vara två företeelser som undersöks för att metoden ska bli komparativ.[22] Vad gäller kvalitativ komparativ analys så betyder det att företeelser jämförs och kausal komplexitet utgör en väsentlig del, det innebär att samma utfall kan ges av olika anledningar där teoretikernas resonemang kan leda till liknande slutsatser fast av olika orsaker.[23] I uppsatsen är den komparativa delen av metoden utmärkande genom att teoretikers resonemang och slutsatser jämförs och analyseras, vilket sker med utgångspunkten med kausal komplexitet i åtanke. Analyserna som dras kommer följaktligen jämföras utifrån uppsatsens teoretiska utgångspunkt med nämnda definition av artdiskriminering och utgår från ovan nämnda definition av moraliskt ansvar.

Metoddiskussion

En kvalitativ metod lämpar sig väl för att möta uppsatsens syfte och anspråk. Detta främst då kvalitativ metod har att göra med människosyn och främst förståelse främst insikt vilket uppsatsen ämnar undersöka. Uppsatsen avser även människans relation till andra individer och även till sig själv då den visar på vad som anses mänskligt. Detta gör att den kvalitativa metoden även fungerar väl i detta avseende då den i hermeneutisk kontext inte bara berör tekniska eller teoretiska aspekter utan även är ett uttryck för människans sätt att förhålla sig till sig själv och världen där människan befinner sig.[24] Detta gör att en kvalitativ metod är väl lämpad för att möta uppsatsens syfte och möjliggör ingångar till metaetiska och filosofiska reflektioner och analyser där problematisering kring vad människor är och var gränsen går mellan människor och andra individer går vad gäller moraliskt ansvar.

Vad gäller komparativa metoder så återfinns stort värde i dessa när det kommer till att undersöka företeelser som förutsägs eller förklaras.[25] Detta möter väl uppsatsens syftesbeskrivning då den avser ett förutsatt framtidsscenario där AI tar större plats i diskussioner kring etik och samhällskonstruktion samtidigt som uppsatsen ämnar förklara olika teoretikers ställningstagande gällande AI utifrån nämnda teoretiska ansats. Komparativa metoder som används kvalitativt kan vidare bidra med analys mellan samspel som kan bero på olika orsaker även om utfallet blir snarlikt, så kallad *kausal komplexitet*.[26] I uppsatsen märks detta genom att kvalitativ analys möjliggör samspel mellan teoretikernas

[22] Denk, Thomas. *Komparativa analysmetoder*. Lund: Studentlitteratur AB. 2012, 11.
[23] Ibid., 63–64.
[24] Starrin, B. Svensson, P–G, 62–63.
[25] Denk, T. *Komparativ metod–förståelse genom jämförelse*, 28.
[26] Denk, T. *Komparativ metod–förståelse genom jämförelse*, 63.

resonemang och visa på flera vägar till samma åsikt och hur de yttrar sig utifrån den teoretiska ansatsen.

6. Disposition

Tidigare har uppsatsens syfte, frågeställningar, materialpresentation, teoretisk anknytning och metod presenterats. I följande kapitel sker en presentation av rådande forskningsläge kring etik och teknik rörande uppsatsens syfte. Sedan följer en bakgrundsbeskrivning av moraliskt ansvar, etik och moral samt människan och teknik som ingång till uppsatsens resultatredovisning. I uppsatsens resultatredovisning lyfts författarnas ståndpunkter för moraliskt ansvar gällande utveckling av artificiell intelligens, potentiella moraliska dilemman samt rättigheter för enskilda AI-individer. Efter resultatredovisningen analyseras det som framkommit för att sedan lyftas till en didaktisk och metareflektion.

7. Forskningsläge

Robotetik?

Robot ethics 2.0 är en bok som innehåller kapitel skrivna av olika författare och forskare som sammanställdes och publicerades av forskaren Patrick Lin 2017. I boken presenteras behovet av att diskutera kulturella och etiska aspekter av teknisk utveckling av robotar. Författarna menar att boken kan verka som en guide för samhället, att erbjudas insikt i hur samhället kan komma att påverkas av robotar samtidigt som den kan fungera omvänt och bidra till hur robottekniker kan förstå hur samhället kommer motta robotar. Boken tar även upp eventuell problematik och presenterar eventuella nya problemområden vad gäller mötet mellan samhället och robotteknik. Boken har bidragit med tankegods och incitament för att ge insikt i hur artificiell intelligens kan komma att påverka samhället vad gäller moraliskt ansvar. Boken framlägger också ett potentiellt scenario med flertalet individer som inte är människor.[27] Bokens ingångsvinkel med etiska aspekter och samverkan mellan människor och robotar har bidragit med tankegods till utformandet av uppsatsens syftesformulering.

[27] Lin, Patrick. *Robot ethics* 2.0. Oxford: Oxford University Press, 2017.

Forskning om AI – teknik

Futurologen Ray Kurzweil skrev boken *Singularity is near: When humans transcend biology* vilken utgavs 2006 där han redogör för forskningsläget under tiden för bokens skrivande samt förutspår teknisk utveckling vad gäller AI som skildras som evidensbaserad. Kurzweil skildrar i boken om sin teoretiska utgångspunkt som han kallar acceleration av lagen om snabbare avkastning som han skildrar som exponentiell teknisk utveckling. Denna utveckling kommer enligt honom nå teknologisk singularitet, vilket innebär den punkt när maskiner blir medvetna och självförbättrande. Kurzweil skildrar hur han anser att tendenser inom forskningen pekar på att punkten för självförbättrande och medveten AI kommer ske runt 2045. Det som enligt Kurzweil gör att denna framtidsvision kommer bli sanning är mått och grafer av teknisk utveckling som Moores lag, vilken mäter antalet transistorer som får plats på ett chip vilken utvecklas i en exponentiell kurva. Utvecklingen kommer innebära att gränsen mellan teknik och mänsklighet suddas ut för att sedan helt sammanföras då AI kommer överskrida mänskligheten samtidigt som människor kommer att kunna förbättras med AI-teknik.[28] Boken har använts som inspiration till uppsatsens undersökning och utgjort en del i bakgrundsbeskrivningen. Kurzweils uppfattning om medveten AI ligger även till grund för det hypotetiska scenario om medveten AI som undersöks utifrån moraliskt ansvar i undersökningen.

Forskning om AI – etik

Vad gäller etiska frågeställningar gällande AI så har Kurzweil även skrivit boken *How to create a mind: When computers exceed human intelligence* där han redogör för hur han uppfattar aktiviteten i den mänskliga hjärnan som en hierarkisk verksamhet med igenkänning av olika mönster. Kurzweil beskriver hur AI–individer kommer kunna bli bärare av konstgjorda hjärnor som fungerar på liknande sätt som biologiska mänskliga hjärnor.[29] Bokens beskrivningar av konstgjorda hjärnor där medvetenhet kan konstrueras artificiellt ligger till grund för det förmodade antagandet om framtida AI-individers preferenser som berörs av moraliskt ansvar. Det är följaktligen medvetenheten som beskrivs av Kurzweil som undersöks av källorna utifrån den teoretiska ansatsen om moraliskt ansvar och artdiskriminering.

8. Bakgrund

[28] Kurzweil, Ray. *Singularity is near: When humans transcend biology*. London: Penguin books, 2006.
[29] Kurzweil, Ray. *How to create a mind: When computers exceed human intelligence*. London: Duckworth overlook, 2014.

Etik och moraliskt ansvar

Huruvida en handling anses bra eller dålig samt vad som avgör under vilken av beskrivningarna en handling betraktas kan anses vara en av världshistoriens tidiga frågeställningar. Sokrates, som var en av de tidigast kända moralfilosoferna, menade att moralen förstås inte som en bagatell utan om hur man bör leva. Detta uttalande kan ses som en inledning till att förstå moral och ansvar för handlingar. Moralfilosofi värderar inte ett konkret, mätsäkert utfall utan inbegriper motstridiga tankar och utgångspunkter medan det inom andra forskningsfält som fysik återfinns en rad etablerade sanningar som vetenskapen vilar på.[30]

Som ovan nämnt återfinns flertalet olika riktningar för att närma sig etiska ställningstaganden och moraliskt ansvar, detta skiljer inte enbart mellan intellektuella och filosofiska olikheter utan kan även anses till viss del kulturellt betingat.[31] En normativ etisk teori om moraliskt ansvar som återfinns bland flertalet filosofer som Ayn Rand är den etiska egoismen. Denna uppfattning menar att människans moraliska ansvar enbart gäller människan självt och då främst på individnivå. Varje enskild människa har alltså främst ansvar för sig själv vilket dock kan innebära handlingar som hjälper andra då det hjälper en själv. Att uppnå sin egen lycka är enligt ovan nämnda Ayn Rand människans högsta syfte och bör betraktas som paradigm inom denna moralfilosofiska strömning.[32] En annan ingång till att förstå moraliskt ansvar kommer från förespråkare av kontraktsetiken där teoretiker menar att moraliskt ansvar är lösningen på ett praktiskt problem, individer samarbetar för gemensam nytta och därför ses handlingar som moraliskt bra eller dåliga. Moraliska regler blir således önskvärda om de ger fördelar för ett gemensamt socialt levnadssätt.[33]

En annan normativ etisk teori som berör ansvar är konsekvensetiken, eller utilitarismen, som avser att maximera mängden goda konsekvenser. Förespråkaren Jeremy Bentham framhåller att moralen syftar till att maximera lyckan i världen. Han menar att det viktigaste moraliska ställningstagandet är nyttoprincipen som innebär att maximera nyttan

[30] Rachels, James. Rachels, Stuart. *Rätt och fel–introduktion till moralfilosofi*. Lund: Studentlitteratur AB. 2011, 7–8.
[31] Ibid., 25–26.
[32] Ibid., 79.
[33] Rachels. Rachels, 99.

med en handling. Det är följaktligen konsekvenserna, snarare än intentionen, som avgör värdet av en handling.[34]

Pliktetiken menar att värdet av en handling kan härledas ur rationellt tänkande, detta yttrar sig genom plikter inför olika handlingar. Dessa plikter utgår alltid från det kategoriska imperativet, som introducerats av filosofen Immanuel Kant, vilket innebär att en handling är god om den kan accepteras som allmän lag och detta bör rationella agenter handla efter. Det är exempelvis alltid fel att ljuga om detta är en handling som en rationell agent betraktar som fel i ett sammanhang.[35] Dygdetiken är en ytterligare normativ etisk teori som menar att det som ger gott är karaktärsdrag som gör att människan agerar moraliskt korrekt. Denna teori avser egenskaper snarare än handlingar och menar att egenskaperna kan odlas och följas till en rimlig nivå vilket leder till vad som anses gott. Dygdernas motsats är laster vilket utmärks genom att de karaktärsdragen inte ger fördelar för de inblandade vilket dygderna gör.[36] Kontraktsetiken förespråkas av moralfilosofen John Rawls som menar att sociala kontrakt mellan individer uppstår för att det gynnar deltagarna och således värderas handlingarna moraliskt utifrån samverkan mellan individerna. Deltagarna i det sociala kontraktet känner till en början inte till varandra väl utan befinner sig bakom en slöja av okunnighet, de kan dock enligt Rawls enas om två rättviseprinciper. Den första är att alla deltagare skall erhålla så stora fri- och rättigheter att det inte krockar med andras likadana fri- och rättigheter. Den andra principen innebär att skillnader deltagare emellan accepteras om de är uppnåeliga för alla och gagnar även de missgynnade deltagarna. Dessa principer menar Rawls ligger till grund för utformandet av sociala kontrakt vilka avgör moraliskt värde utifrån dess betydelse för deltagarna.[37]

Människans ansvar gentemot andra

Vad gäller etiska teorier så svarar de på frågor om vad som generar gott, detta gör att moraliskt ansvar att handla i den riktning teorin företräder förespråkas. Detta har traditionellt främst avsett människans moraliska ansvar gentemot andra människor även om det sedan antiken funnits diskussion kring moraliskt ansvar gentemot andra livsformer. På senare tid kan ansvar även anses beröra naturen självt, detta har föranletts av en artegoism där nya kunskaper också bidragit med insikt i att moraliskt ansvar för naturen

[34] Ibid., 117.
[35] Ibid., 161–163.
[36] Ibid., 187–189.
[37] Rawls, John. *En teori om rättvisa*. Göteborg: Daidalos AB. 1999, 55–76.

även gynnar mänsklighetens fortlevnad och välbefinnande.[38] Moraliskt ansvar gäller alltså inte enbart relationen människor emellan utan även andra former av liv, även om ansvar inte utkrävs så är det ett ställningstagande vad gäller relationen mellan arter eller individer.

Flertalet filosofer använder sig av normativa etiska teorier för att argumentera gällande relationen mellan människa och djur. Peter Singer exempelvis argumenterar utifrån utilitarismen att det är orätt att orsaka onödigt lidande oavsett arttillhörighet då det genererar negativa konsekvenser.[39] Även de som motsätter sig Singers slutledning behöver ändå förhålla sig till relationen mellan livsformer och att applicera moraliska frågeställningar som moraliskt ansvar gentemot andra arter är ett test i konsekvens gällande etiska teorier. Även filosofen Tom Regan skriver om moraliskt ansvar i form av rättigheter för icke-mänskliga djur vilket han menar att vi genom rationellt tänkande kan förstå. Regan menar att vi har plikt att ge vissa rättigheter till djur då de sannolikt försetts med liknande egenskaper som ligger till grund för rättigheter som skyddar människor. Exempelvis har människor och djur sannolikt en önskan om att undvika onödigt lidande vilket gör att rättigheter enligt Regan bör upprättas för att förhindra detta.[40] Även Regans användande av etiska teorier appliceras gällande andra individer vilket testar konsekvens.

Etiska frågeställningar berör moraliskt ansvar, en som skriver om detta är filosofen Thomas M. Scanlon vilken skriver om ansvar gentemot andra. Han menar att vad som gör att en handling betraktas god eller inte beror på omdömen som rör andra individer. Scanlon resonerar kontraktualistiskt och menar att handlingar är goda beroende på dess värde gentemot andra.[41] Frågor som rör moraliskt ansvar är att betrakta som huruvida en handling kan tillskrivas en aktör på så sätt att det krävs för att kunna utgöra en grund för moralisk värdering, något som Scanlon kallar ansvar som tillskrivenhet. Detta innebär att en person är ansvarig för en given handling då det är befogat att utföra denna som då ligger till grund för moralisk värdering av agenten som utfört handlingen. Enligt Scanlon kan även ansvar vara användbart i olika sammanhang gällande olika roller, en individs ansvar inbegriper förpliktelser och anspråk gentemot andra och är beroende av valmöjligheter gällande att välja olika beslut.[42] Relationen mellan individer framstår således som central vad gäller moraliskt ansvar och utfallet gentemot andra och sig själv avgör följaktligen värdet av en

[38] Jonas, 214–215.
[39] Singer, 6–7.
[40] Regan, 7–18.
[41] Scanlon, Thomas. M.*Vad är vi skyldiga varandra*. Göteborg: Daidalos AB, 10–15.
[42] Scanlon, 241–242.

handling. Enligt Scanlon betraktas handlingar som moraliska då de ligger till grund för restriktioner som vi tillsammans accepterar för att erhålla skydd gentemot skadligt beteende från andra eller sig själv.[43]

Människan och teknik

Människan kan anses ha utvecklats tekniskt under hela vår existens, ända från mänsklighetens bemästrande av enkla verktyg och eld till nu har vår teknologiska förmåga ökat och den utgör en stor del av vår vardag. Den tekniska utvecklingen har ständigt använts för att underlätta människans vardag och i modern tid har teknologin gjort att välståndet väl överskrider vad som krävs för ett materiellt ombonat liv.[44]

Hur artificiell intelligens relaterar till människor kan, enligt teoretikern Thomas B. Kane och dennes teori om artificiell intelligens innebära ett förmodat scenario och förhållningssätt då artificiell intelligens är ett verktyg för människan likt många andra teknologiska uppfinningar även om AI innebär ett nytt paradigm då de även kan betraktas som artificiella personer. Detta innebär enligt Kane att AI behöver problematiseras vad gäller etik.[45]

För att göra det möjligt att undersöka uppsatsens material utifrån den teoretiska ansatsen krävs ett hypotetiskt scenario om när AI kommer kunna replikera mänskliga karaktärsdrag, något som ännu inte är fullt möjlig eller i varje fall inte allmänt gods. Därför behövs en bedömning om självbestämmande och medvetna AI–individer som tillåts en subjektiv upplevelse av verkligheten. Kurzweil menar att detta uppnås år 2029, han menar att detta år har AI och nanoteknik samverkat på ett sådant sätt att konstgjorda hjärnor gör det möjligt att replikera medvetande och subjektiv upplevelse bland artificiellt intelligenta robotar.[46] Även om denna förutsägelse ligger i framtiden och kan framstå som tveksam så är det en liknande framtidsbild som uppsatsen ämnar undersöka, för att applicera moraliskt ansvar gentemot AI krävs att de har kapacitet att vara någon och inte enbart en icke-medveten dator. Det framgår enligt Kurzweil att hans uppfattning om världen år 2029 kommer befolkas av robotar och AI som deltar i det samhälleliga livet på ett sätt som enbart människor gör i skrivande stund, AI kommer enligt honom exempelvis att ta plats i

[43] Ibid., 258–259.
[44] Von Wright. Georg–Henrik. *Vetenskapen och förnuftet.* Stockholm: Albert Bonniers förlag. 2003, 130–131.
[45] Kane, Thomas. B. *A framework for exploring intelligent artificial personhood*. Edinburgh: Napier university, 255–257.
[46] Kurzweil. *Singualarity is near*, 220–221.

utbildning, rehabilitering, kommunikation, sjukvård, politik och så vidare. Kurzweil skriver om hur världens befolkning kommer bestå av mänsklig och icke-mänskliga individer som erhåller liknande ansvar i civilisationen. Trots att datorer och AI förmodas kunna klara turingtestet, vilket innebär att de är intelligenta nog att ersätta människor i konversationer, menar han att det finns politiska och filosofiska diskussioner om AI och människors rättigheter. AI:s subjektiva upplevelser förmodas vara allmänt accepterade även om ökad intelligens bland maskiner menas sudda ut gränsen mellan människa och maskin där ansvar för framtida utveckling av civiliserat liv inte självklart kan enbart avse människor.[47] Framtidsscenariot som presenteras anses särskilt relevant då den tekniska utvecklingen skildras vara i behov av moralisk och filosofisk slutledning vad gäller människor och icke-människor.

Den teknologiska utvecklingen gällande AI har potential att överskrida mänskligt förstånd, detta är något som Kurzweil framhåller. Han menar att från 2029 och framåt kommer artificiellt intelligenta robotar delta i samhällslivet och erhålla intelligens och medvetenhet som väl replikerar den mänskliga biologiska hjärnan. Dessa individer kallar han nanorobotar och hjärnans funktioner som leder till subjektiva upplevelser som rädsla, njutning och ilska kommer enligt honom nu att kunna skapas av robotar genom sammankoppling av artificiella nervceller. Han menar dessutom att mänskliga sinnen kan laddas upp i artificiella motsvarigheter till hjärnor vilket liknas vid en överföring av ens medvetenhet och gränsen mellan människa och maskin blir således otydligare.[48] Det är denna förutsägelse som används i uppsatsens resultatredovisning där AI klarar turingtest och har motsvarigheter till mänskliga egenskaper, karaktärsdrag och medvetenhet. Dessa AI-individer placeras i källornas resonemang för att möta uppsatsens syfte om moraliskt ansvar.

9. Resultatredovisning

Som ovan nämnt framgår det att det är högst troligt att självmedveten AI kommer utvecklas i framtiden. Det är Kurzweils definition av medveten AI som undersöks nedan utifrån moraliskt ansvar. Vidare kommer även moraliskt ansvar för utvecklingen av AI undersökas och dessa synsätt sedan undersökas mot Singers definition av artdiskriminering.

[47] Ibid., 222–225.
[48] Ibid., 313–315.

Författarna resonerar i olika riktningar vad gäller moraliskt ansvar, vissa skriver däremot mer om ansvar för enskilda individer, utveckling av AI samt risker och möjligheter med AI.

Företrädare för att AI inte ska utvecklas
Paul Conrad Samuelsson

Teoretikern Paul Conrad Samuelsson argumenterar gällande moraliskt ansvar för enskilda AI-individer men även om utveckling och risker. Han menar att artificiellt medvetande troligtvis kommer kunna vara kännande och att utvecklingen av dessa AI-individer kommer ta plats i samhället inom en snar framtid. Han skildrar i sin artikel, som ovan nämnts, de risker han menar kommer återfinnas vid en sådan utveckling, Samuelsson menar att teknisk utveckling av AI innebär risker och att tekniska framsteg inte är en garant för moraliska framsteg utan att det kan vara det motsatta.[49] Samuelsson framstår som tveksam till teknologisk utveckling där AI blir kännande individer och skulle kunna vara att betrakta som personer och innehavare av egenskaper som vad gäller människor skyddas av mänskliga rättigheter och lagar vilka härleds ur moraliskt ansvar.

Samuelsson resonerar utilitaristiskt gällande moraliskt ansvar mellan människor och AI, detta yttrar sig genom att han menar att det finns två stora moraliska problem med utvecklandet av AI. Människans ansvar är således enligt Samuelsson att varsamt kontrollera teknologisk utveckling där moral menas vara en tongivande orsak. Samuelsson lutar sitt resonemang mot vad han kallar teknokulturella ikoners varnande ord och han menar att:

> ”De varnar alla för risken att en självförbättrande AI med ett dåligt definierat mål kommer att döda alla människor för att uppfylla målet, ungefär med samma känslomässiga inblandning som vi förstör myrstackar när vi bygger motorvägar”.[50]

Samuelsson grundar sin argumentation i två spår där den första är att människans moraliska ansvar att hålla tillbaka utveckling av AI är rätt med hänsyn till lidande bland AI. Samuelsson hänvisar till möjliga konsekvenser och menar att utveckling av AI kan göra att dessa individer kan lida i oändlig mängd med hög kognitiv upplevelse.[51] Han hävdar

[49] Samuelsson, 33.
[50] Samuelsson, 33–34.
[51] Ibid.

alltså att människans moraliska ansvar ligger i att inte utveckla AI som kan uppleva evigt lidande och att konsekvenserna, det vill säga lidande, blir enligt honom negativa.

Samuelsson problematiserar dessutom utilitaristiskt resonemang i motsatt inriktning, han menar då att utilitarismen skulle kunna användas omvänt med hög kognitiv förmåga hos AI som skulle möjliggöra evigt välmående. Han menar att detta är ett felaktigt resonemang då lidande och välmående inte är motsatser utan välmående avtar över tid och lidandet har enligt Samuelsson en liknande effekt även om gränsen går högre. Detta innebär att evigt lidande är värre än vad välmående är positivt och han menar därför att det etiska förhållningssättet gentemot AI bör fokuseras på minimering av lidande snarare än på maximering av lycka med hänvisning till kvalitativa skillnader i lidande och välmående.[52] Samuelssons uppfattning är att vi inte vet vad konsekvenserna blir vad gäller utveckling av AI. Han menar att detta också borde vara en indikation på att utveckling bör ske varsamt:

> ”Stående bakom okunnighetens slöja, skulle du vara redo att satsa, även med goda odds, på en värld där du kanske kommer utsättas för den största möjliga mängden tortyr, för alltid?”[53]

Samuelssons andra spår är att utveckling av AI riskerar att skapa negativa konsekvenser även för människor och världen i stort, han menar att teknologiska bakslag riskerar att bli förödande för allt levande, människor inkluderat. Enligt Samuelsson har teknisk utveckling alltid inneburit bakslag där kärnkraftsolyckor utgör exempel i hans resonemang som dessutom åsyftar att teknologisk utveckling av AI kommer leda till existentiella risker för hela planeten vilket han menar kan härledas ur programmeringsfel eller felaktigt användande av AI-teknik.[54] Samuelssons resonemang har således två dimensioner där potentiellt lidande för AI och annat liv tillsammans med existentiella risker för alla är resonemang till att teknologisk utveckling ska hållas tillbaka.

Enligt Samuelsson är följaktligen det huvudsakliga moraliska ansvaret gällande AI från mänskligt håll att hålla tillbaka utvecklingen vilket han menar gynnar både AI och människor samt världen i allmänhet. Främst då riskerna trumfar möjligheterna vad gäller lidande respektive välmående. De huvudsakliga anledningarna till att moraliskt ansvar blir att hålla tillbaka utvecklingen härleds ur Samuelssons tankar om att den teknologiska utvecklingen sker betydligt mycket snabbare än väsentliga mänskliga aspekter som moral, kultur och politik. Moraliskt ansvar för utveckling av AI ses som omoralisk på

[52] Ibid., 35.
[53] Ibid., 36.
[54] Samuelsson, 41.

grund av riskerna enligt Samuelsson som dessutom menar att utvecklingen bör förbjudas och det vore önskvärt om ingen vore intresserad av teknologisk utveckling. Avslutningsvis skildrar Samuelsson hur människans hybris innebär att vi tror oss vara berättigade till allt som vi anstränger oss för och på så vis ställer mänskligheten sig bortom moralen.

> ”Vår kroniska oförsiktighet och obryddhet gör artificiella medvetanden till vår största etiska risk.”[55]

Vad gäller moraliskt ansvar framgår det genom hans resonemang att Samuelsson varnar för utvecklingen och att AI medför stora risker både vad gäller AI och övriga individer samt världen i stort. Hans resonemang utgår från utilitarismen där han anser att konsekvenserna riskerar att bli mer negativa än positiva samtidigt som han menar att lidande är värre än välmående är positivt. Samuelsson tar även upp beskrivet problem om att teknisk och social utveckling inte sker i liknande takt utan att det finns problem i att den tekniska utvecklingen sker så pass snabbt att övrig utveckling inte lyckas följa och moralisk mognad vad gäller de etiska problem som kan återfinnas med teknologi riskerar att inte uppnås.

Bill Hibbard

Bill Hibbard beskriver främst moraliskt ansvar i hur AI kommer bli moraliska agenter men även hur de kan komma att verka i samhället. Enligt Hibbard finns risker med utveckling av AI även om han anser att det är troligt att AI kommer återfinnas som aktörer i framtiden, han menar att etiska problem finns i att se till att AI kan navigera och göra val som är etiska. Problemen ligger i att få AI att räkna ut vad som är moraliskt korrekt. Hibbard menar att detta innebär att människor har moraliskt ansvar att utveckla modellbaserade verktyg för att föra samman förmåga att räkna ut moraliskt korrekta handlingar med omgivningen för att undvika kontextbaserade moraliska övertramp.[56] Det moraliska ansvaret kan alltså härledas till att ge AI rätt möjligheter att agera moraliskt i olika situationer. Hibbard använder sig av John Rawls rättviseteori för att argumentera för hur AI bör fungera i samhället. Han beskriver hur Rawls första princip, där individer har rätt att maximera grundläggande friheter, mycket väl kan appliceras även för AI liksom Rawls andra princip om att ojämlikheter bör gynna de mindre bemedlade för att anses rättfärdiga. Hibbard menar att Rawls teori fungerar bättre vad gäller AI än

[55] Ibid., 42.
[56] Hibbard, 76.

exempelvis utilitarismen som syftar till att maximera goda konsekvenser, han anser att de värden som gör att konsekvenser kan anses goda är synnerligen subjektiva och kan anses vara endemiska för människan. Enligt Hibbard gör detta att utilitarismen inte passar väl när det kommer till att inkludera AI i ett samhälle med mänskliga aktörer.[57] Det framstår också att moraliskt ansvar ligger i att inkludera AI i sociala kontrakt där rättighetsprinciperna inbegriper även dem och på så sätt skyddas AI från moraliska felsteg.

Vidare tar Hibbard upp etik rörande att testa AI för utveckling. Han menar att utveckling och testning kräver insyn och rationalitet där AI bör fungera på ett önskvärt sätt för människor snarare än tvärtom. De stora utmaningarna med detta menar han är intressekonflikter vid utveckling av AI där olika aktörer strävar efter olika mål, detta löses genom öppenhet där allmänheten har tillgång till resonemangen som förs gällande utveckling och aktörernas målsättningar.[58] Moralisk ansvarsfullhet återfinns här i att se till att utvecklingsprocessen av AI sker på ett moraliskt godtagbart sätt. Avslutningsvis menar Hibbard att AI har stor potential men också medför stora risker, han menar att då AI ger stora möjligheter att hjälpa människor bör vetenskapligt fokus läggas vid utvecklingen samtidigt som AI innebär stora risker för människan bör den vetenskapliga utvecklingen kompletteras av väl underbyggda etiska slutsatser. Detta menar han kommer, tillsammans med god insyn, bidra till finansiering av etisk AI som kan ha positiv inverkan snarare än negativ.[59] Av Hibbards slutsatser framgår det att människans huvudsakliga moraliska ansvar är att se till att utvecklingen sker på ett korrekt sätt, först då kan AI se som etiskt. Och innan detta är fastställt, finns inte moraliska incitament för att utveckla AI-teknologi.

Michael R. LaChat

Michael R. LaChat skildrar moraliskt ansvar för utvecklingen snarare än enskilda individer och han tar även upp metaetiska ställningstaganden gällande personlighet och intelligens. Enligt LaChat behandlas individer som erhåller artificiellt medvetenhet som bland annat klarar ett turing-test som något han kallar personlig AI.[60] Han menar att det är skillnad på intelligens och personlighet där intelligens skildras som icke-

[57] Ibid., 87.
[58] Ibid., 127.
[59] Ibid., 142.
[60] LaChat, 70.

standardiserad respons på yttre stimulans medan personlighet är något som finns utanför intelligens, ett slags sinne medan intelligens kan representeras av en hjärna eller något liknande. Detta kallar LaChat ontologisk dualism och han menar att personlig intelligens är ett problem för utvecklandet av AI:

> "A personal intelligence must have personality, and this seems on the face of it to be an almost impossible problem for AI."

Det är alltså enligt LaChat viktigt med en emotionell upplevelse för att intelligens ska bli personlig, han menar också att en rädsla för döden ligger till grund för varandet.[61] Vad gäller människans moraliska ansvar gentemot AI menar LaChat att utvecklandet av en intelligent personlig AI bör följa strikta riktlinjer eftersom det medför större risker än förmåner i dagsläget. Han tar också upp skillnaden mellan AI och människor i att mänsklig reproduktion historiskt har varit nödvändigt för mänsklighetens överlevnad medan utvecklandet av AI för närvarande innebär ett lyxproblem där dock det positiva utfallet skildras som att ge medvetet liv vilket han dock menar även sker vid reproduktion. Vad gäller reproduktion så menar han dessutom att förhållandet mellan risker och fördelar är bättre för tillfället än vad gäller AI och drar slutsatsen att utvecklandet av AI som personlig bör vägas mot eventuella risker vilket i dagsläget ses som omoraliskt.[62]

Vidare menar LaChat att det inte är självklart att en personlig AI bör erhållas moralisk hänsyn i den mån FN:s deklaration om mänskliga rättigheter föreskriver. Han menar exempelvis att det finns en motstridighet i att förbjuda robotar att hamna i slaveri när mänskligheten utvecklat teknik i detta syfte. Han menar dock att frågeställningar kring rättigheter anses orealistiska då AI kommer byggas i lager vilket gör att moraliskt ansvar kring dem kommer lösas i takt med utvecklingen. LaChat resonerar även kring hur moraliskt ansvar kan härledas ur fri vilja:

> "If free will is real in some sense, there is again no reason to believe that it might not be an emergent property of a sophisticated level of technical organization, just as it might be asserted to arise through a slow maturation process in humans. I should also add that not all AI-experts are convinced an AI could not attain free will."

[61] Ibid., 72–73.
[62] Ibid., 75.

Han ser alltså att personlig AI skulle kunna hållas moraliskt ansvarig för sina handlingar om deras handlingar kan härledas ur fri vilja.[63] Förutom personlighet så menar LaChat att känslor är en viktig faktor för att omfattas av moraliskt ansvar, han anser alltså att AI behöver känslor och en intelligent anledning för att kunna omfattas av moraliskt ansvar. Han använder sig av en observatörsteori för att styrka detta påstående där allvetenhet, allmännytta och empati utgör grunden för moralisk värdering. Avslutningsvis menar LaChat att utvecklingen av AI med stor sannolikhet kommer fortsätta vilket gör att det är viktigt att vara medveten om etiska risker eftersom det är människans försök att skapa något mänskligt och personligt:

> "Much caution and forethought are necessary when we contemplate the human construction of the personal."

Han fastslår också att AI kan vara nästa steg i evolutionen vilket gör att ett visst moraliskt ansvar för AI återfinns, om än utvecklat i en eftertänksam process, samtidigt som det kan skapa nästa generations människa. LaChat menar att det finns två tydliga positioner vad gäller moraliskt ansvar för AI där den ena är ensidigt förespråkande och den andra är avskrivande, han menar att det är moraliskt korrekt att låta utvecklingen ske under vissa premisser där hans uppfattning om etikens första premiss är att inte göra skada. Denna premiss menar han bör prägla utvecklingen av AI och han avslutar med att beskriva att han känner sig exalterad inför en sådan utveckling.[64] LaChat tar inte tydligt ställning för eller emot vilket genomsyrar hans resonemang, det framgår istället att han menar att moraliskt ansvar avser vissa ramar och att AI kan omfattas av moralisk hänsyn om den kan bevisas ha känslor, intelligens, fri vilja och har mer fördelar än nackdelar, alltså en personlig AI som kan bevisas vara innehavare av dessa. En sådan AI skulle enligt honom både vara moraliskt ansvarig för sina handlingar och omfattas av moralisk hänsyn om än på människans villkor även om det framgår att han anser att detta kan vara svårt att säkerställa. Vad gäller människans moraliska ansvar att utveckla AI tar han positionen att det är moraliskt under kontrollerade former även om vissa moraliska oklarheter ännu återfinns.

Delanalys

[63] LaChat, 75–76.
[64] Ibid., 76–79.

Samuelssons position utgår från människans behov men även risker för AI, han framstår som skeptisk till teknisk utveckling utan moralisk reflektion då det går att skönja genom hans resonemang om moraliska risker vad gäller. Samuelsson använder utilitarismen som stöd för sin position där han framhåller att konsekvenserna, för både människor och AI, riskerar att bli negativa snarare än positiva. Både Hibbard och Samuelsson använder sig av normativa etiska teorier för att framföra sina argument där de kommer fram till liknande slutsatser fast på olika grunder från Rawls rättviseteori och utilitarismen. LaChats hållning ses som något mer positiv till utveckling av AI men han bedömer dock riskerna som större än möjligheterna i dagsläget vilket gör att vissa moraliska frågeställningar bör klargöras innan utveckling kan tillåtas. LaChats resonemang anses delas med tidigare nämnda Hibbards där försiktighet är ledord som kan utläsas ur deras resonemang.

Förespråkare av försiktig utveckling av AI
Stefan Larsson och Anna Felländer

Stefan Larsson och Anna Felländer skriver om moraliskt ansvar i utvecklingen av AI och hur samhället kan komma att påverkas eller inte av AI-individer. Enligt Larsson återfinns det huvudsakliga mänskliga ansvaret vad gäller AI i att skapa en hållbar AI. Han tar upp att AI tenderar att anamma diskriminering som förekommer i samhället som exempelvis olika behandling utifrån kön, ålder, klass, hudfärg och så vidare.[65] Anna Felländer tar i samma rapport upp moraliska dilemman som kan uppstå med självlärande AI, hon menar att sättet AI samlar information om människor eller andra individer hamnar i konflikt med de berördas integritet. Hon tar också upp problem att använda medveten självlärande AI och sjukvård där kommunikation mellan vårdtagare och AI vilket kan drabba sjukvården negativt. Just inom vården återfinns i och med patienternas självbestämmande och teknologisk innovation riskerar att hamna i värdekonflikt, hon menar att det finns osäkerhet kring vilka värden som bör premieras. Felländer drar slutsatsen att det behövs ett paradigmskifte vid utveckling av AI och maskininlärning om artificiellt medvetna individer ska ta plats på arbetsmarknaden.[66] Det framgår att enligt Larsson och Felländer återfinns risker i det moraliska ansvaret då det riskerar att drabba enskilda människor i verksamheter som ännu inte anpassats till en tid då AI deltar i samhället.

[65] Larsson, 13–14.
[66] Felländer, 34–36.

För att möta ovan nämnda utmaningar föreslås hållbar AI av Larsson och Felländer, detta innebär att utvecklingen av AI präglas av rekommendationer:

> ”Hållbar AI kräver att vi
> 1. fokuserar regleringsfrågor i vid
> mening,
> 2. stimulerar mångvetenskap och samverkan, samt att
> 3. tillitsbyggande i användningen av samhällsapplicerad artificiell
> intelligens och maskininlärning är centralt och kräver mer
> kunskap i relationen mellan transparens och ansvar.”

De tenderar alltså en varsam utveckling av AI där dessa regleringsfrågor präglar processen vilket gör det moraliskt ansvarsfullt och på så sätt möts de etiska och sociala utmaningarna med AI.[67] Det framgår att det huvudsakliga ansvaret från mänskligt håll består i att utveckla en hållbar AI snarare än rättigheter eller moraliskt ansvar gentemot AI-individer. Det anses framgå att Felländer och Larsson värderar det moraliska ansvaret gentemot vårdtagare och andra människor där AI kan tänkas verka bör värderas först vid utveckling av AI snarare än rättigheter för artificiellt medvetna individer eller moralisk skyldighet i att utveckla AI för goda syften.

Eliezer Yudkowsky

Yudkowsky skildrar främst utveckling av AI snarare än om rättigheter för enskilda individer och menar att de största riskerna med AI är att människan avslutar forskningen för tidigt för att på riktigt förstå möjligheterna. Han förutser att de risker som kan återfinnas med utveckling av AI kommer ur partiskhet för människor och mot andra kognitiva agenter.[68] För att undvika negativa konsekvenser förespråkar Yudkowsky utveckling av vänlig AI vilket skildras som ett moraliskt påbud, en vänlig AI innebär enligt honom att dessa AI-individer kan implementera mänskliga värden när de väger olika handlingsalternativ. En superintelligent AI kan enligt honom dock ändra sina målsättningar men Yudkowsky menar att han anser positiva scenarier som mer sannolika och dessutom kan möjliggöra stor mängd välmående och ökad livskvalitet vilket gör att insatserna motiverar utvecklingen.[69]

[67] Ibid., 44.
[68] Yudkowsky, 1–2.
[69] Yudkowsky, 12–13.

Det framgår ytterligare att Yudkowsky ser AI som en möjlighet snarare än en risk vad gäller etik då han medger att det är omöjligt att förutse hur välvillig eller illvillig AI kan komma att påverka världen. Dock verkar det som att AI möjliggör utveckling som ger positiva utfall på sikt även om vissa bakslag kommer på vägen. Han tar upp exempel på annan historisk utveckling som att människor på stenåldern knappast kunnat föreställa sig samtida samhällskonstruktioner medan de flesta skulle anse att utvecklingen gett positivt utfall. Yudkowsky menar att detsamma gäller människans nuvarande position och framtida utveckling där AI kan komma att utgöra en viktig komponent.[70] Det moraliska ansvaret för AI kan alltså förstås som att inte bromsa utvecklingen och inte förse AI med kapacitet att handla moraliskt korrekt. Detta verkar han mena kunna förverkligas med utveckling av vänlig AI.

Yudkowsky beskriver också hur vänlig AI kan skydda mänskligheten mot illvillig AI, forskare kommer också enligt honom att lära av tidigare misstag och korrigera den tekniska utvecklingen så att utfallet blir successivt bättre vilket gör att AI handlar mer i linje med vad människan föredrar.[71] Han menar att utmaningarna med AI kan mötas med precist utformande:

> "We must execute the creation of Artificial Intelligence as the exact application of an exact art. And maybe then we can win."

Enligt honom har AI potential som bör hanteras med försiktighet för bästa utfall och detta görs möjligt med vänlig AI som mål för utvecklingen.[72] Det huvudsakliga moraliska ansvaret kan utläsas bestå i utvecklandet av en vänlig AI och inte låta vår oro stå i vägen för den teknologiska utvecklingen vilken har möjlighet att bidra till moraliskt önskvärda utfall där möjligheterna trumfar riskerna.

Jack Copeland

Copeland skildrar moraliskt ansvar för utveckling av AI från ett individuellt perspektiv och problematiserar även på metaetisk nivå kring medvetenhet. Han beskriver hur AI för diskussionen om människor är maskiner till sin spets. Han skriver att människans beståndsdelar fungerar likt maskiners och han tar dessutom upp att utvecklingen av AI bidragit med kunskap om hur fri vilja och medvetna handlingar kan uppstå fysiskt och mekaniskt. Copeland menar också att det inte finns något som indikerar att även

[70] Ibid., 25.
[71] Ibid., 27–29.
[72] Ibid., 43.

självmedvetenhet och kännande inte skulle vara fysiska reaktioner.[73] Även om det inte skrivs ut kan dock Copelands position tolkas som att det inte finns så mycket som skiljer människor och AI åt vilket skulle indikera moraliskt ansvar för AI som för människor. Även hans beskrivning av hur utvecklingen av AI har lärt människan mer om filosofiska frågeställningar kring fri vilja och medvetenhet kan tolkas som att moraliskt ansvar ligger i att tillåta utveckling av AI.

Copeland skriver också att 1600-tals filosofen René Descartes yttrande: jag tänker, alltså är jag kan kompletteras för att bli: jag tänker, alltså är jag en maskin vilket gör uppfattningen mer anpassad till 2000-talet.[74] Denna hållning cementerar Coplands hållning om AI vilket indikerar att AI och människor har långt fler likheter än skillnader vilket således ligger till grund för hur moraliskt ansvar appliceras.

Max Tegmark

Max Tegmark argumenterar främst kring utveckling av AI som individer samt även om rättigheter för enskilda individer. Enligt fysikern Tegmark är det sannolikt att AI i framtiden kommer kunna förstå och tillämpa olika målformuleringar för att sedan självständigt utföra dem, hur detta värderas menar han avgörs av människans moraliska ansvar gentemot AI.[75] Tegmark beskriver att det moraliska ansvaret måste utredas innan superintelligens utvecklas, han hävdar att det på långa vägar inte råder någon konsensus om vilken etisk hållning som är att föredra men att det finns fyra principer som han menar är universella. Dessa principer är maximering av positiva, och minimering av negativa upplevelser, mångfald av positiva upplevelser är att föredra, medvetna individer bör erhålla självstyre och förening av uppskattade scenarion och avståndstagande från oönskade scenarier. Tegmark menar att den första principen, som han liknar vid utilitarism, bör yttra sig i människans moraliska ansvar genom att ge möjlighet för maximalt goda konsekvenser samtidigt som de präglas av andra principens antagande om mångfald bland upplevelser vilket han anser mer lyckobringande än monoton lycka. Tegmark skriver även att principen om självstyre utifrån individens preferenser bör ligga till grund för utformandet av AI-individer, likt FN:s deklaration om mänskliga rättigheter fast utifrån AI-individers önskemål. Tegmark anser att dessa principer hör samman då självstyre maximerar goda konsekvenser och bidrar till mångfald av positiva upplevelser.

[73] Copeland, 249.
[74] Ibid., 249.
[75] Tegmark, 358.

Enligt Tegmark innebär den fjärde principen att AI bör ha rätt att påverka framtida kontext vilket kan tillsammans med självstyre ses som demokratiska ideal.[76]

Tegmark belyser även moraliska dilemman som riskerar att uppstå om AI enbart brukas för människan, förhållandet dem emellan bör istället präglas av ömsesidighet vilket han menar kan härledas ur två rättigheter som inbegriper ovan nämnda principer, den första är att en medveten individ har frihet att tänka, lära, kommunicera, äga egendom och får inte skadas eller förstöras. Den andra rättigheten innebär att en medveten individ har full frihet samt rättighet så länge det inte strider mot första rättigheten. Tegmark anser dock inte att denna princip är fullständig utan att moraliska dilemman kan uppstå där skydd av svaga vägs mot rättigheter för starka individer.[77] Det är alltså enligt Tegmark inte självklart hur rättigheter för AI bör utformas utan detta är något som i dagslägets moraliska diskussion riskerar att inbegripa motsägelser, det moraliska ansvaret gentemot AI tenderar att ligga i utveckling av hållbar etisk ställning innan den tekniska utvecklingen kan implementeras i samhället fullt ut. Tegmark likställer dock AI och människor vad gäller hållare av rättigheter utan att mena att de skulle vara endemiska för människan.

Avslutningsvis menar Tegmark att det inte finns universella uppfattningar om hur AI ska behandlas etiskt vilket gör att framtiden bör präglas av omfattande moraliska diskussioner kring hur detta bör hanteras i takt med den tekniska utvecklingen. Han menar att mänskligheten bör göra detta innan den tekniken försett AI med medvetenhet och superintelligens varpå rättigheter för AI kan realiseras. Tegmark ser dock inte detta som något hinder utan han menar att strävan efter en perfekt etisk hållning riskerar att hålla tillbaka utvecklingen av huvudsakligen goda upplevelser. Han menar att vissa etiska ståndpunkter är självskrivna vilka med fördel redan nu skulle kunna implementeras i AI-individer.[78] Tegmarks hållning präglas av försiktighet samtidigt som han menar att AI har stor etisk potential, det framgår att hans tydligaste åsikt om moraliskt ansvar från mänskligt håll avser att människan bör komma överens om en gemensam etik. Detta bör alltså leda till konsensus trots sociokulturella skillnader mellan olika grupper av människor. Det verkar enligt Tegmark som att människans utarbetande av en gemensam

[76] Tegmark, 361–363.
[77] Ibid., 364.
[78] Ibid., 365.

hållning är den största moraliska problematiken och moraliskt ansvar kan således utläsas i utvecklingen av en sådan gemensam etik.

Mathias Risse

En förespråkare av vissa rättigheter och filosofisk diskussion kring AI är Mathias Risse, han menar att medvetenheten förändrar sättet att betrakta mänskliga rättigheter vad gäller maskiner. Detta kan anses logiskt då medvetenhet och upplevelse spelar en avgörande roll i formandet av mänskliga rättigheter. Enligt Risse kan utvecklingen mot individer som liknar människor, även de artificiellt skapade, göra att ansvar att behandla dessa väl successivt ökar i takt med den tekniska utvecklingen. Risse skriver följande om likheter mellan varelser och medvetenhet:

> "If the mind just is a complex algorithm, then we may eventually have little choice but to grant the same moral status to certain machines that humans have. Questions about the moral status of animals arise because of the many continuities between humans and other species: the less we can see them as different from us in terms of morally relevant properties, the more we must treat them as fellow travelers in a shared life."[79]

Det framgår att Risse menar att utvecklandet av AI skiljer sig från moralisk samverkan mellan andra livsformer och människor då människor haft kontroll över dem och kunnat skapa forum för åtskillnad mellan andra livsformer och människor och därefter format deklarationen om mänskliga rättigheter. Detta ska enligt Risse dock inte neka moralisk hänsyn till icke-mänskliga djur men lämnar stort individuellt skydd för människor medan utveckling av AI snarare kan mynna ut i ett omvänt scenario där AI har möjlighet att ansvara för oss snarare än tvärtom. Detta gör enligt Risse att mänskligheten bör utveckla AI på ett sätt att mänskliga rättigheter respekteras och även rättigheter mot andra artificiella individer även om de har förmåga att ignorera dem och agera i motsatt riktning:

> "They would have to be designed so they respect human rights even though they would be smart and powerful enough to violate them. At the same time they would have to be endowed with proper protection themselves. It is not impossible that, eventually, the UDHR would have to apply to some of them."[80]

Mänskliga rättigheter kan alltså i framtiden komma att omfatta andra livsformer, samtidigt som ansvar mellan icke-mänskliga också skildras som viktigt. Risse finner

[79] Risse, 3.
[80] Ibid., 8.

dock viss problematik med den teknologiska utvecklingen då värdejustering bör implementeras i en snabbt utvecklande AI som arbetar i företag och politiska organisationer, dessa områden kan komma att stå i konflikt vad gäller värdeförpliktelser. Företag som arbetar i linje med mänskliga rättigheter förväntas respektera detta medan AI som arbetar i sådana företag riskerar ha olika förpliktelser som kan stå i kontrast till varandra där de strävar efter vinst för företaget samtidigt som värdeuppdraget åberopas. Detta blir följaktligen en utmaning att implementera i samverkan med den teknologiska utvecklingen.[81]

Risse beskriver även problematik om mänskliga rättigheter som garant för positivt utfall av moraliskt ansvar, han tar upp att AI likaväl kan komma att behöva implementera människor i deras rättighetsuppfattning och om de vill förlänga moraliskt ansvar till även människor.[82] Framtidens AI kan komma att förändra sättet att se på rättigheter i grunden och Risse menar att den kommer att sätta punkt för upplysningstidens ideal om rättigheter och moraliskt ansvar, detta för att artificiell intelligens som överskrider människans möjligen inte ämnar följa mänskliga rättigheter eller helt enkelt väljer ett annat paradigm för vad som anses skapa gott. Det kan vara ett förhållningssätt som övergår människans föreställningar och härleds ur andra företeelser än upplysningstidens tankeströmningar. Avslutningsvis menar Risse att den största risken som återfinns bland utveckling av AI är den sociala ojämlikhet som riskerar uppstå i takt med teknologins framväxt. Han menar att dessa två, social ojämlikhet och utveckling av AI i samverkan, riskerar att prägla de kommande 70 åren och utmana rådande föreställningar om mänskliga rättigheter.[83]

Virginia Dignum

Virginia Dignum resonerar främst kring moraliskt ansvar i att utveckla eller inte utveckla AI och menar att den tekniska utvecklingens hastighet gör att AI numera bör diskuteras utifrån juridiska, politiska och etiska aspekter. Hon framhåller att AI ofta framställs som en komponent i en dystopisk utveckling där världen blir sämre vilket enligt henne inte är en rättvis bild av AI. Dignum framhåller att AI idag redan hjälper många människor inom sjukvård, transport, service och säkerhet vilket en ytterligare självmedveten AI med stor sannolikhet också skulle göra. Eftersom självbestämmandet och medvetenheten är en av

[81] Risse, 9.
[82] Ibid., 10.
[83] Ibid., 15.

de delar av AI som utvecklas i hög hastighet anser Dignum att ansvar är något centralt som behöver problematiseras vad gäller AI och andra livsformer. Hon menar att AI skapas i en värld som präglas av mänskliga ideal vilket gör att AI behöver utrustas för att verka i en sådan värld, därför ligger människans huvudsakliga moraliska ansvar i att förbereda ramverk för etiskt handlande för AI:

> ”These frameworks must deal both with the autonomic reasoning of the machine about such issues that we consider to have ethical impact, but most importantly, we need frameworks to guide design choices, to regulate the reaches of AI systems, to ensure proper data stewardship, and to help individuals determine their own involvement.”[84]

Det återfinns enligt henne alltså ett moraliskt ansvar i att ge AI verktyg att verka etiskt. Det framgår dessutom att Dignum inte ser AI som en etisk risk utan snarare möjlighet, en möjlighet som dock har vissa utmaningar att handskas med innan de kan verka med maximalt positivt resultat. Vidare menar Dignum att de etiska värdena som vi bör utrusta AI med är beroende av sociokulturella sammanhang där mänsklighetens ansvar ligger i att skapa system för hur AI kan navigera mellan samhällen, kulturer och förklara sina resonemang för att garantera insyn och skapa förtroende. Dignum föreslår därför en utveckling av så kallad ansvarig AI där sådana system eller individer tillåts verka likt människor, alltså förklara sig, resonera och hållas ansvariga för sina handlingar. För att detta ska kunna implementeras menar hon att människan har ansvar att utveckla intelligenta system för grundläggande mänskliga principer. Detta kan enligt henne garantera människans fortsatta trygghet och utveckling i en hållbar värld.[85]

Enligt Dignum är följaktligen autonomiskt ansvar vägen framåt för utvecklingen av ansvarig AI. Hon skildrar hur AI relaterar till AI på olika plan:

> “Ethics by Design: the technical/algorithmic integration of ethical reasoning capabilities as part of the behaviour of artificial autonomous system;
>
> Ethics in Design: the regulatory and engineering methods that support the analysis and evaluation of the ethical implications of AI systems as these integrate or replace traditional social structures;

[84] Dignum, 1.
[85] Ibid., 1–2. .

> Ethics for Design: the codes of conduct, standards and certification processes that ensure the integrity of developers and users as they research, design, construct, employ and manage artificial intelligent systems."[86]

Vad gäller algoritmer och beteende föreslår Dignum programmering av ett socialt kontrakt som korrekta mekanismer för styrning och reglering av beteende. Visionen för ett sådant innebär att algoritmerna som påverkar olika individer måste vara transparenta, rättvisa och ansvariga utifrån värden som delas av berörda parter. Detta kräver också att ansvarig AI kan felsöka och kompensera samt övervaka avvikelser från etiskt godtagbara handlingar.[87] Även om AI bör hållas ansvarig tar hon upp att det finns ett behov i att säkerställa att ansvarig AI verkar för mänskliga intressen och garanterar navigering i potentiellt motstridiga sammanhang. Dignum tar även upp resonemang om moraliskt ansvar att kunna stoppa AI vid avvikelser eller farligt beteende. Detta beskrivs kunna yttra sig genom ett nödstopp där mänskliga operatörer kan avleda eller avbryta en handling. Denna säkerhetsmekanism skildras även som en riskfaktor då ansvarig och medveten AI skulle kunna bli medveten om möjligheten att avsluta sin existens vilket belyser behovet av självkorrigering. Slutligen belyser hon även vikten av utvecklandet av en scenariogenereringsmekanisk som gör det möjligt att simulera olika handlingar och värdera dem vilket maximerar möjligheterna för positivt utfall och minimerar risker. Ett sådant system skulle enligt Dignum också verka säkrare än ett nödstopp som är beroende av mänskliga faktorer.[88] Det huvudsakliga moraliska ansvaret mellan människor och AI verkar enligt Dignum vara att utveckla AI som kan verka etiskt och agera i en värld styrd av mänskliga teorier och värderingar. Människans ansvar ligger följaktligen i att förse AI med verktyg för att agera etiskt och således förse AI med potential att vara moraliskt ansvariga för sina handlingar och ansvara för mänskligt välbefinnande.

Nick Bostrom

Bostrom skriver om AI som individer men främst kring hur dessa kan komma att verka i samhället men även om moraliska risker och möjligheter. Det framstår som att AI-individer kommer delta i samhällslivet och erhålla en intelligens och kompetens som överskrider människans något som Nick Bostrom kallar superintelligens, han menar att moraliska dilemman uppstår vad gäller rättigheter i arbetsliv och han framhåller risker i att AI blir slavar trots att de kan vara självmedvetna. Bostrom menar att AI kan betraktas

[86] Dignum.
[87] Ibid., 2.
[88] Ibid., 2–3.

som lönearbetare anställda av människor vilket kan vara att föredra över slaveri-liknande tillstånd där AI ägs av människor. Det framstår inte som självklart hur utvecklingen kan fortlöpa enligt Bostrom utan han menar att det kan bli så att AI utvecklas som volontärarbetande fria agenter som tycker om att bidra, dessa individer kan erhålla lön och värdiga rättigheter som härleds ur socioekonomiska mallar för välmående.[89] Vad gäller arbetsmarknaden och samhällslivet framgår det alltså att Bostrom förespråkar ansvarsfull kontroll där utfallet premieras, det är alltså moraliskt ansvarsfullt att låta AI ta plats efter dessa premisser.

Bostrom menar också att superintelligent AI kan komma att skapa medvetna och kännande simulationer, möjligtvis i syfte att interagera med människor och verka i samhället styrd av mänskliga preferenser. Han menar att dessa simulationer med stor sannolikhet skulle komma att förstöras när de uppfyllt sitt syfte, likt djur inom djurförsök, vilket skulle innebära stora moraliska dilemman. Bostrom menar att om dessa simulationer innehåller kännande och självmedvetna varelser, som människor, vilka vid experimentets slut upphör att existera riskerar utfallet bli ett folkmord som är moraliskt fel. Han tar också upp att superintelligenta AI-individer kan upprepa dessa simulationer i triljoner scenarion vilket skulle innebära evig massutrotning. Bostroms resonemang för varför möjliggörande av dessa simulationer är moraliskt klandervärdiga är att riskerna är så pass stor mängd lidande för en stor mängd individer vilket enligt honom inte skulle väga upp utfallet av att superintelligent AI lär sig interagera med människor. Dessutom menar han också att det finns praktiska moraliska problem inbyggda i förmågan att simulera kännande individer, om AI kan göra detta kan de efter egna intressen behandla dessa individer utan moralisk hänsyn och hålla dem gisslan.[90] Även om Bostrom i detta fall inte drar någon slutsats om hur moraliska dilemman som ovan beskrivna bör lösas, kan det utläsas att moraliskt ansvar ligger i att värdera dessa risker innan utvecklingen av superintelligent AI.

Andra risker som Bostrom beskriver är att AI inte nödvändigtvis har ett beteende som liknar människans och dessutom är svårt att utreda. Detta menar han kan ställa till stora besvär om AI erhåller positioner i samhället som polis och rättsväsende. Han menar att detta är en risk vid utveckling av AI som självständiga agenter vilka istället bör präglas av människoliknande karaktärsdrag då de kan verka på ett sätt som ger mindre moraliska

[89] Bostrom. *Superintelligence: Paths, dangers and strategies*, 167–168.
[90] Ibid., 126.

risker och bättre utfall.[91] Ytterligare vad gäller moraliskt ansvar skriver Bostrom att människan bör förse AI med målsättningen att göra moraliskt gott, detta skulle förstås ur AI-individernas höga kognitiva förmåga. Han menar att en sådan utveckling, kallad moralisk riktighet, kan förse AI med verktygen att agera moraliskt korrekt även bättre än människor då kognitiv förmåga och intelligens skulle leda AI till mer rationella och logiska beslut. En sådan moralisk kompass, installerad i AI, skulle minimera riskerna för moraliska felsteg även om de inte är till nytta för eller till och med hotar mänskligheten. En sådan utveckling skulle enligt Bostrom innebära att vi får acceptera risker för oss själva för moralisk riktighet. Han påtalar att det finns risker kring uppfattningar om vad moralisk riktighet är, det kan vara kontext och/eller individbundet, detta menar Bostrom kan överkommas genom superintelligens vilket gör att AI kan navigera mellan moraliska kontexter och göra riktiga slutledningar.[92]

Nick Bostrom skriver även att utveckling av AI är väl motiverat då potentialen bland AI-individer är betydligt större än hos människor som begränsas biologiskt. Denna potential menar han finns inte bara inom intelligens utan även inom moraliskt tänkande.[93] Moraliskt ansvar för innebär alltså att tillåta utveckling av mer kapabla och moraliska individer än oss människor. Bostrom menar att den stora utmaningen ligger i att utrusta AI med människovänliga preferenser. Något som kan komma att krocka med AI:s eget moraliska ställningstagande vilket inte nödvändigtvis betyder att främja människan.[94] För att ta sig an utmaningen om människovänlig AI krävs en vänlig utgångspunkt, moraliskt ansvar innebär alltså att människan bör utrusta AI med målformuleringar som är moraliskt korrekta. Bostrom menar att detta kan leda till en vänskaplig, hållbar relation mellan människa och maskin:

> ”A “friend” who seeks to transform himself into somebody who wants to hurt you, is not your friend. A true friend, one who really cares about you, also seeks the continuation of his caring for you.”

Vänskapliga relationer till människor tryggar på så sätt mot moraliska felsteg.[95] Det framgår alltså att AI kan betraktas som vänliga och förhållandevis jämlika människan, Bostrom beskriver också att moraliskt ansvar föreligger huruvida vi bör utveckla AI eller

[91] Bostrom. *Superintelligence: Paths, dangers and strategies*, 202–206.
[92] Ibid., 215–218.
[93] Bostrom. *Ethical issues in advanced artificial intelligence*, 1.
[94] Ibid., 1–2.
[95] Ibid., 4.

inte. De problem han identifierar är att det finns risker att AI används i destruktiva syften och att vi inte skulle vara redo att ta oss an en sådan utmaning medan möjligheterna med sådan AI är botemedel mot sjukdomar, ökat välstånd och evigt liv med hjälp av sinnesöverföring. Bostrom tar ställning för möjligheterna snarare än mot riskerna och menar att så snart det bedöms möjligt och säkert bör utvecklingen ske för att möjliggöra ovanstående önskvärda utfall.[96] Bostroms position vad gäller moraliskt ansvar blir alltså att AI bör utvecklas då möjligheterna trumfar riskerna vilket gör det till ett moraliskt påbud. Han skildrar inte att AI bör erhålla särskilda rättigheter men att de kan betraktas som mänsklighetens vänner och på så vis utvecklas till en ömsesidig partner snarare än underordnad.

Delanalys

Författarna ovan kan ses hålla en försiktig men positiv hållning till AI där de redogör för risker men menar att möjligheterna överträffar dem och att utveckla AI är alltså moraliskt ansvarsfullt. De skriver inte något konkret om rättigheter för enskilda AI individer som rättighetshållare utan utgår från ett mänskligt perspektiv när de resonerar och de flesta gör hypotetiska antaganden om hur framtiden med AI kommer se ut. Sammanfattningsvis kan författarnas position beskrivas som att AI ses som något övervägande gott och möjligheterna gör det moraliskt ansvarsfullt att låta AI utvecklas. Dignum, Bostrom och Risse resonerar utifrån att AI bör utrustas med verktyg att agera i en mänsklig värld medan Tegmark resonerar utilitaristiskt och kommer fram till att AI troligtvis kommer innebära ett positivt utfall av konsekvenser snarare än negativt.

Förespråkare för rättigheter och utveckling av AI
Dorna Behdadi

Dorna Behdadi skriver om AI som individer och hur dessa kan drabba människan men även hur människan kan drabba AI-individer. Hon resonerar även på metaetisk nivå kring medvetenhet. Enligt Behdadi finns ett moraliskt dilemma vad gäller medveten AI, hon härleder detta ur det faktum att medvetenhet är svår att bevisa deduktivt där alla individers medvetenhet inte är sammankopplad med andras. Hon menar att medvetenhet är nära sammankopplat med moraliskt ansvar, om en individ är att betrakta som medveten bör de också omfattas av moraliskt ansvar.

[96] Bostrom. *Ethical issues in advanced artificial intelligence*, 5.

> ” There is something it is like to be them.”[97]

Det är således moraliskt att ansvara för medvetna individer vilka har en subjektiv upplevelse av verkligheten enligt Behdadi. Medvetna individer skildras som medlemmar i ett moraliskt sammanhang som inte kan behandlas efter en agents tycke. Hon tar upp att många varnar för den teknologiska utvecklingen av AI då den hypotetiskt skulle kunna utgöra ett hot mot människor vilket beskrivs som en ensidig hållning då den bortser från medvetande. Då medvetenhet enbart kan antas menar hon att detta bör vara praxis för moraliskt ansvar, detta motiveras med att vi låter information som kan observeras ligga till grund för antaganden om medvetenhet. Denna information kan vara kommunikation, rörelse eller även mätbar information som aktivitet i hjärnan. Vad gäller AI menar Behdadi att dessa individer kan testas för att avgöra medvetenhet genom att låta dem interagera med en människa och om denna människa inte kan avgöra om individen den samverkar med är människa eller AI bör AI betraktas som medvetet i samma grad som människor då beteendet indikerar detta. Behdadi beskriver att vissa skulle mena att AI bara skulle kunna härma medvetet beteende utan att vara medveten själv, det är enligt henne ett dilemma då vi härleder medvetenhet till observerbar information som likväl kan vara härmad.[98]

Vidare menar Behdadi att vi människor måste förberedas för framtida AI som tar plats i samhället genom att synliggöra partiskhet för människor och mot AI. Hon härleder detta från människans historia då partiskhet varit grunden för omoralisk aktivitet gentemot andra människor eller icke-mänskliga djur genom krig, slaveri och utrotning av andra arter. Detta gör att medveten AI som är skapad att fylla mänskliga behov innebär ett moraliskt dilemma eftersom framtida medveten AI kan hamna i en utsatt situation om de nekas moralisk hänsyn. Hon menar att frågan om hot är felställd:

> ”I do not think that they would only pose a risk to us, we might pose a much greater risk to them.”[99]

Hon menar avslutningsvis att om en individ visar de rätta indikationerna på medvetenhet borde deras härkomst inte spela någon roll utan fokus om moraliskt ansvar bör istället ligga på likheterna, alltså observerbar information om medvetenhet. Sådana individers

[97] Behdadi.
[98] Ibid.
[99] Ibid.

existens förtjänar moralisk hänsyn enligt Behdadi även om hon uttrycker osäkerhet kring huruvida människan är kapabel att ta sig an detta moraliska dilemma.[100]

Delanalys

Behdadi använder inte bara ett mänskligt perspektiv utan utgår från AI när hon resonerar. Hon gör dessutom deduktiva antaganden om att individer beter sig på ett sätt som indikerar att de är kännande så finns det god anledning att tro att individerna är det. Detta kan liknas vid naturrättsteoretiskt tänkande och hon härleder sedan detta till att rättigheter borde erhållas för AI. Behdadis position avskriver partiskhet vilket kan leda till ogrundad diskriminering, avslutningsvis visar hon dock på en viss skepticism kring utvecklingen då vi människor kanske inte är moraliskt förberedda.

10. Analys och slutdiskussion

Hur omfattas artificiell intelligens av vårt moraliska ansvar enligt teoretiker som resonerar om ämnet?

Det framgår att författarna menar att moraliskt ansvar ligger i att om AI bör utvecklas så ska detta ske på ett säkert sätt som inte riskerar moraliska felsteg, vare sig det gäller relationen mellan AI och människor eller tvärtom. Det är också tydligt att de anser att utveckling av AI är ett moraliskt ansvar för människan som medför möjligheter men även risker vilket är de fundament som olika författare resonerar utifrån. Flertalet författare menar dessutom att det är människans moraliska ansvar att se till att AI erhåller moraliska verktyg att verka i ett samhälle konstruerad av människor samtidigt som ett par författare tillskriver AI rättigheter.

Vilka likheter respektive skillnader finns mellan de olika författarna?

De huvudsakliga skillnaderna består i att vissa författare inte anser att AI bör utvecklas då de anser de etiska riskerna som större och existentiella än möjligheterna som kan uppstå. Några författare tar istället en försiktigt positiv position där de menar att utveckling av AI bör ske då möjligheterna till välmående för AI och människor ökar, detta menar de dock kräver försiktighet då felsteg kan innebära existentiella risker för människan. Ett par författare tillskriver också AI rättigheter för att skydda deras välmående och existens. Vissa

[100] Behdadi.

författare använder sig av olika teorier som utilitarism och kontraktsetiken för att styrka sin position medan andra resonerar praktiskt.

Vad kan dessa likheter respektive skillnader antas bero på?

Likheterna och skillnaderna anses bero på författarnas olika utgångspunkter där några har bakgrund inom vetenskaplig forskning kring AI, några är etiker och en är futurolog. Detta gör att de är bekanta med olika forskningsfält och närmar sig problemet från olika positioner. Även synen på mänskligheten antas även ligga till grund för likheter och skillnader, vissa författare är mer måna om att säkerställa människans existens medan andra har en mer objektiv ingång.

Hur kan slutsatserna liknas vid Peter Singers teori om artdiskriminering?

Det framgår att om AI utvecklas på så sätt som beskrivs i bakgrunden där de erhåller karaktärsdrag som är jämförbara med människan men inte skyddas av moralisk hänsyn eller rättigheter skulle det innebära en form av diskriminering som kan liknas vid artdiskriminering mellan människor och djur. Det som skiljer är att människan inte skapat djuren medan AI är att betrakta som en del i mänsklighetens tekniska utveckling. Dock kan det sägas att människan definitivt påverkat tama djur efter eget tycke och kontinuerligt skapar nya individer.

Analys

I resultatredovisningen framgår det klart att Det finns två huvudpositioner vad gäller moraliskt ansvar där den ena är att varsam utveckling av AI är moraliskt korrekt medan den andra innebär att människan inte är redo för moraliska och existentiella risker vilket gör att utvecklingen bör hållas tillbaka. Utöver dessa positioner finns två ytterligare åsikter där den mest AI-positiva tar ställning för att AI bör utvecklas och är ett moraliskt påbud, dessutom framhåller företrädare av denna hållning upprättande av rättigheter för medvetna AI-individer. Den motsatta hållningen menar istället att det är moraliskt ansvarsfullt att inte utveckla AI och att en sådan utveckling har betydligt större risker än möjligheter, företrädare för denna position hänvisar till moralisk omognad att handskas med moraliska felsteg som riskerar att uppstå. Författarna resonerar på olika grund och utgår från olika scenarion, främst Samuelsson och Behdadi, men även Risse, diskuterar gällande moraliskt ansvar för individer medan de flesta övriga författarna skildrar moraliskt ansvar i att utveckla eller inte utveckla artificiell intelligens. Detta antas bero på vilken typ av forskning de olika författarna bedriver där Behdadi och Samuelsson har en mer

hermeneutisk vinkel och övriga forskar inom positivistiska och naturvetenskapliga områden som fysik och futurologi. Flertalet författare skildrar främst moraliskt ansvar utifrån hypotetiska scenarion där främst människans välbefinnande utgör utgångspunkten.

Vad gäller förespråkare av rättigheter för AI förs resonemang för upprättande mot bakgrund av deduktiva antaganden om AI-individers förmåga att uppleva en subjektiv verklighet och lidande. Denna hållning kan väl appliceras även vad gäller människans förhållande till icke-mänskliga djur och drag av artdiskriminering återfinns således inte. Vad gäller andra resonemang om moraliskt ansvar som försiktig utveckling eller avståndstagande från att utveckla AI överhuvudtaget finns en antropocentrisk tendens där människans välbefinnande skildras som centralt och det som gör att talespersoner för dessa positioner kan anses resonera på ett sätt som liknas vid artdiskriminering. För att denna position skulle vara konsekvent skulle dessa personer avsäga samma utveckling om mänskligheten var i den position som AI-individer skulle befinna sig i. Om författarna menar att människor inte skulle utvecklas då moraliska felsteg skulle vara möjliga och existentiella risker för ansvariga aktörer även skulle användas som resonemang mot utveckling av mänskligheten. För att undvika att resonemangen förs godtyckligt och inbegripa artdiskriminering skulle de följaktligen behöva acceptera dessa hypotetiska resonemang. Vad gäller Tom Regans definition av artdiskriminering innebär den att avsäga livssubjekt som rimligen innehar egenskaper som ligger till grund för rättigheter som innehas av människor. AI-individer är att betrakta som livssubjekt och nekande av rättigheter dem liknas väl vid hans definition av begreppet. Regan hänvisar även till evolutionen vad gäller medvetenhet, han skriver exempelvis att naturen troligtvis försett andra livssubjekt än människan med medvetenhet. Vad gäller AI kan det anses vara en produkt av människan snarare än evolutionen även om människan givetvis är en produkt av evolutionen. Detta påverkar dock inte rationella antaganden om medvetenhet som Regan beskriver, AI kan liksom andra livssubjekt som agerar på ett sätt som indikerar att de vore medvetna så är de troligtvis det.[101] Vad gäller Peter Singers definition av artdiskriminering innebär den att kännande individer inte bör utsättas för lidande då konsekvenserna blir oönskade. Även här kan AI appliceras, att neka kännande och medveten AI moralisk hänsyn skulle följaktligen innebära artdiskriminering. Singer skriver om jämlikhetsprincipen vilket väl inbegriper AI där den innebär lika hänsyn efter

[101] Regan, 45–46.

de inblandades preferenser.[102] Även detta kan användas för att förklara antidiskriminering om inte att inte ge lika hänsyn till inblandades intressen och AI kan rimligen anses vara individer som drabbas av detta. De författare som i resultatredovisningen avskriver moralisk hänsyn för AI kan anses tendera artdiskriminering medan om de ställer sig kritiska till utveckling kan detta anses moraliskt om konsekvenserna för de inblandade överväger de negativa. De olika författarna i resultatredovisningen resonerar på olika sätt för att företräda sin hållning, de flesta argumenterar utifrån praktiska risker och möjligheter medan andra använder sig av normativa etiska teorier för att avgöra huruvida utveckling av AI och tillskrivande av rättigheter för AI är moraliskt påbud.

Intressant är att Samuelsson och Tegmark båda lutar sina ståndpunkter mot utilitaristisk teori om att maximera goda, och minimera onda, konsekvenser men de kommer fram till olika slutsatser. Avskrivande av rättigheter för existerande AI kan liknas vid artdiskriminering, oavsett om den anses moralisk av företrädare, medan en kritisk hållning till utvecklande av AI inte nödvändigtvis behöver innebära artdiskriminering. Detta eftersom ännu icke-existerande individer inte kan anses hållare av rättigheter samtidigt som författarna framhåller att AI medför etiska möjligheter.

Slutdiskussion

Att studera moraliskt ansvar och AI anses bidra med nya plattformar för diskussioner om moraliska dilemman samtidigt som det bidrar med nya infallsvinklar på metanivå där diskussionen om AI ytterligare bidrar med nya frågeställningar kring vad mänskligt egentligen innebär. Vad gäller AI så ger en diskussion om moraliskt ansvar nya tankegångar kring hur människan bör förhålla sig till andra individer där AI utmärker sig då det är en annan livsform som har möjlighet att nå högre intelligens än människan. Dessa nya tankegångar anses väl avvägda då de även kan bidra med tankegods för arbete om moral och etik i grundskolan och gymnasieskolan. Moraliskt ansvar gentemot AI kan ge elever en ytterligare dimension vilket hjälper dem att skapa förståelse för tillämpad etik vad gäller moraliska dilemman, normativa etiska teoriers synsätt på hur AI behandlas samt etik på metanivå där AI fungerar som en faktor i vad som anses mänskligt. Artificiell intelligens kan betraktas som ett fält som ännu ligger början av sin utveckling, både vad gäller teknisk och etisk forskning. Detta märks av i hur författarna skildrar moraliskt ansvar. Vad gäller moraliskt ansvar från både naturvetenskapliga och hermeneutiska fält

[102] Singer, 31–33.

gör detta att undersökningen erhållit flertalet dimensioner och detta kan meds fördel genomföras i andra undersökningar. Förhoppningsvis kan fortsatt diskussion om moraliskt ansvar bidra till utvecklingen av människans förståelse för etik och möjligtvis förbereda oss bättre på att hantera moraliska dilemman som riskerar att uppstå med annan teknisk utveckling. Den tekniska utvecklingen av AI sker i skrivande stund i hög hastighet och ytterligare undersökningar av etik och AI kan således föra teknik och etik närmare varandra.

11. Käll- och litteraturförteckning

Litteratur

Denk, Thomas. *Komparativa analysmetoder*. Lund: Studentlitteratur AB. 2012.

Denk, Thomas. *Komparativ metod–förståelse genom jämförelse.* Lund: Studentlitteratur AB. 2002.

Jonas, Hans. *Ansvarets etik–utkast till en etik för den teknologiska civilisationen*. Göteborg: Bokförlaget Daidalos AB. 1991.

Kane, Thomas. B. *A framework for exploring intelligent artificial personhood*. Edinburgh: Napier University.

Kurzweil, Ray. *Singularity is near: When humans transcend biology.* London: Penguin books, 2006.

Kurzweil, Ray. *How to create a mind: When computers exeed human intelligence*. London: Duckworth overlook, 2014.

Lin, P. *Robot ethics* 2.0. Oxford: Oxford University Press, 2017.

Rachels, James. Rachels, Stuart. *Rätt och fel–introduktion till moralfilosofi*. Lund: Studentlitteratur AB. 2011.

Rawls, John. *En teori om rättvisa*. Göteborg: Daidalos AB. 1999.

Regan, Tom. *Djurens rättigheter–en filosofisk argumentation*. Nora: Nya doxa. 1999.

Scanlon, Thomas. M. *Vad är vi skyldiga varandra*. Göteborg: Daidalos AB.

Singer, Peter. *Djurens frigörelse*. Nora: Bokförlaget Nya doxa. 1999.

Starrin, Bengt. Svensson, Per–Gunnar. *Kvalitativ metod och vetenskapsteori*. Lund: Studentlitteratur AB.2009.

Von Wright. Georg–Henrik. Vetenskapen och förnuftet. Stockholm: Albert Bonniers förlag. 2003.

Källor

Behdadi, Dorna. *Conscious AI: A moral dilemma.* TedxGöteborg. 2019.

Bostron, Nick. *Ethical issues in advanced Artificial intelligence.* Science Fiction and Philosophy: From Time Travel to Superintelligence. Oxford: Oxford University. 2003

Bostrom, Nick. *Superintelligence: Paths, dangers and strategies*. Oxford: Oxford University Press. 2016.

Copeland, Jack. *Artificial Intelligence: A philosophical introduction.* New Jersey*:* Blackwell Publishing. 1993.

Dignum, Virginia. *Ethics in artificial intelligence: introduction to the special issue.* Ehics and information technology. 2018.

Hibbard, Bill. *Ethical artificial intelligence*. Berkeley: Space Science and Engineering Center
University of Wisconsin–Madison
and
Machine Intelligence Research Institute. 2014.

LaChat, Micheal. R. *Artifficial intelligence and ethics: An exercise in the moral imagination*. The AI magazine. Vol 7. Nr 2. 1986.

Larsson, Stefan. Anneroth, Mikael. Felländer, Anna. Felländer-Tsai, Li. Heintz, Fredrik. Cedering Ångström, Rebecka. *Hållbar AI: inventering av kunskapsläget för etiska, sociala och rättsliga utmaningar med artificiell intelligens.* Stockholm: AI sustainability center. 2019.

Risse, *M. Human rights and Artificial intelligense – An urgently needed agenda.* Carr center for human rights policy. Cambridge, Massachusetts. 2018.

Samuelsson, Paul–Conrad. *Artificiella medvetanden är vår största etiska risk*. Filosofisk tidskrift. Stockholm: Thales förlag. 2019.

Tegmark, Max. Liv 3.0: *Att vara människa i den artificiella intelligensens tid*. Stockholm: Volante. 2017.

Yudkowsky, Eleizer.. *Artificial intelligence as a Positive and Negative factor in global risk.* I Global catastrophic risks. Bostrom, N. Crikovic, M. New York: Oxford university press. 2008.