Artificiell Intelligens

- Den osäkra framtiden och riskerna med AI-teknologi.

Kandidatuppsats inom informatik

Ludwig Carlsson
Marcus Malmberg
Wictor Skarp

VT 2023: 2023KANI07


HÖGSKOLAN
I BORÅS

Svensk titel: Artificiell Intelligens: Den osäkra framtiden och riskerna med AI-teknologi.

Engelsk titel: Artificial intelligence: The uncertain future and risks with AI-technology.

Utgivningsår: 2023

Författare: Ludwig Carlsson, Marcus Malmberg och Wictor Skarp

Handledare: Daniel Yar Hamidi

Förord

Vi vill inleda med att skicka ett stort tack till vår handledare Daniel Yar Hamidi som under hela arbetsprocessen hjälpt oss något enormt. Vi vill även tacka respondenterna för att de tog sin tid till en intervju och möjliggjorde denna studie.

Abstract

Artificial intelligence (AI) has received enormous attention recently as one of the most exciting technologies of our time. At the same time, it has also become one of the most debated technologies due to the uncertainty and concerns surrounding its development. This study has been conducted to investigate whether the uncertainty surrounding the development of AI is justified and what risks may arise if development continues without human understanding and control. The study is based on data from qualitative interviews with experts in the subject as well as collected and analyzed articles to gain a broader understanding of the risks and uncertainty associated with the development of AI.

The research questions that the study focuses on answering are:

- Is the uncertainty with AI development justified?
- What risks can arise with AI's continued uncontrolled development without human understanding?

The results from this study shows that the uncertainty surrounding the ongoing AI development is justified, but that it is not something people need to be worried about at the present time as the development is considered to be under control. AI is also not seen as a technology that will or should work completely autonomously, an interaction between the technology and humans will always be required. The result shows that the biggest risk with AI development is that the technology is developed based on the wrong intentions and is used in a harmful way. There is an imminent risk that AI could become a catastrophic force leading to the downfall of humanity, wiping out and replacing humanity. Other risks that could be identified with the help of this research are that AI can generate offensive, incorrect or misleading information. To avoid AI becoming an inhibiting factor for creativity and innovation, it is important to increase knowledge and understanding of AI and ensure its correct use. By ensuring that development goes in the right direction and is compatible with people's goals and values, uncertainty and potential risks are reduced.

The possibilities for future research with the help of this study are vast as AI is a constantly developing field and can be used for many different purposes. From the study's results and conclusions, an imagined five-step model has been developed as a proposal for future research. This five-step model can be applied to analyze and investigate how to manage the uncertainty and risks of AI identified in this research in the future. An additional suggestion that this study proposes as future research is to examine the uncertainty surrounding AI generally in society.

Keywords: Artificial Intelligence, AI, Ethics, Uncertainty, Risks

Sammanfattning

Artificiell intelligens (AI) har fått en enorm uppmärksamhet på senare tid som en av de mest spännande teknologierna i vår tid. Men samtidigt har den också blivit en av de mest omdiskuterade teknologierna på grund av den osäkerhet och oro som omger dess utveckling. Denna studie har genomförts för att undersöka huruvida osäkerheten kring utvecklingen av AI är befogad och vilka risker kan uppkomma om utvecklingen fortsätter utan mänsklig förståelse och kontroll. Studien baseras på data från kvalitativa intervjuer med experter inom ämnet samt insamlade och analyserade artiklar för att få en bredare förståelse för de risker och den osäkerhet som är förknippad med utvecklingen av AI.

Forskningsfrågorna som studien har till fokus att besvara är:

- Är osäkerheten med AI-utvecklingen befogad?
- Vilka risker kan uppkomma med AI:s fortsatta okontrollerade utveckling utan mänsklig förståelse?

Resultatet från denna undersökning visar att osäkerheten kring den stundande AI-utvecklingen är befogad men att det inte är något människor behöver vara oroliga för i dagsläget då utvecklingen anses vara under kontroll. AI ses heller inte som en teknologi som kommer eller bör fungera helt autonomt utan det kommer alltid att krävas en interaktion mellan teknologin och människor. Resultatet visar att den största risken med AI-utvecklingen är att teknologin utvecklas utifrån fel intentioner och används på ett skadligt sätt. Det finns en överhängande risk att AI kan bli en katastrofal kraft som leder till människans undergång genom att utplåna och ersätta mänskligheten. Andra risker som kunnat identifierats med hjälp av denna undersökning är att AI kan generera kränkande, felaktig eller missvisande information. För att undvika att AI blir en hämmande faktor för kreativitet och innovation är det viktigt att öka kunskapen och förståelsen kring AI samt säkerställa dess korrekta användning. Genom att säkerställa att utvecklingen går åt rätt håll och är förenlig med människors mål och värderingar minskar osäkerheten samt de potentiella riskerna.

Möjligheterna till framtida forskning med hjälp av denna undersökning är stora då AI är ett ständigt utvecklande område och kan användas för många olika syften. Från studiens resultat och slutsatser har en tänkt femstegsmodell utvecklats som ett förslag till framtida forskning. Denna femstegsmodell kan tillämpas för att analysera och undersöka hur man i framtiden kan hantera osäkerheten och riskerna med AI som har identifierats i denna undersökning. Ett ytterligare förslag som denna studie föreslår som framtida forskning är att undersöka osäkerheten kring AI generellt i samhället.

Nyckelord: Artificiell Intelligens, AI, Etik, Osäkerheter, Risker

Innehållsförteckning

1. Inledning

1.1 Bakgrund

Under det senaste decenniet har det skett enorma framsteg i datorers prestanda att hantera uppgifter utan mänsklig inblandning. Idag kan datorer hantera och utföra uppgifter som tidigare bara varit möjligt för människor. AI är ett område som är svårt att definiera men handlar om en maskins förmåga att efterlikna människors arbete när det kommer till kommunikation, resonemang och självständigt arbete i bekanta och nya scenarier. Du-Harpur, Watt, Luscombe och Lynch (2020) beskriver AI som en form av maskininlärning eller djupinlärning som handlar om att maskiner kan lära sig genom att hantera märkt träningsdata med hjälp av algoritmer och statistiska modeller. Genom denna process kan maskinen sedan lära sig att känna igen och härleda mönster (Du-Harpur et al. 2020).

AI är idag ett populärt begrepp inom informatik. AI handlar om förmågan hos tekniska system att uppfatta sin omgivning, hantera vad de uppfattar och lösa problem i syfte att uppnå ett specifikt mål. Ökningen av tillgänglig processorkraft och den exponentiella tillväxten och variationen av data har de senaste åren lett till en snabb utveckling av AI. Idag är AI utformat för att lösa väldefinierade problem som filtrera spam, visa snabbaste vägen till en destination eller rekommendera liknande produkter inom e-handel (Techworld, 2021).

Morikawa (2017) undersöker i sin studie om verksamheters attityder och tankar kring ämnet AI. Undersökningen visar att både verksamheter med och utan produktion ser användningen av AI som något positivt. De beskriver att AI kan användas för att förbättra produktiviteten i ett företag vilket kan öka ett företags potentiella tillväxttakt. Detta medför att verksamheter har höga förväntningar på de positiva effekter AI kan generera för deras verksamhet. De höga förväntningarna och den positiva syn verksamheter har på de effekter AI kan medföra stärker författarens argument, att utvecklingen och spridningen av AI kommer ständigt att fortsätta (Morikawa, 2017).

Trots de positiva förväntningarna av AI har experter inom området offentligt väckt en oro kring den snabba utvecklingen i ett brev som signerats av bland annat industrimagnaten Elon Musk och Steve Wozniak. I det öppna brevet *Pause Giant AI Experiments: An Open Letter* beskrivs AI utvecklingen som snabb och potentiellt farlig, vilket tidigare oftast hyllats med lovord och framtidsförhoppningar. AI utvecklingen förklaras vara utom kontroll där utvecklare försöker framställa de mest kraftfulla AI systemen. Den digitala och kraftfulla AI tekniken har nått en punkt där dess skapare inte längre kan förstå, förutsäga eller kontrollera utvecklingen. Frågor som *Ska vi låta maskiner översvämma våra informationskanaler med propaganda och osanning? Ska vi automatisera bort alla jobb, inklusive de som uppfyller dem? Ska vi utveckla icke-mänskliga sinnen som så småningom kan överträffa, överlista och ersätta oss? Ska vi riskera att förlora kontrollen över vår civilisation?* ställs i det öppna brevet. Då ovissheten kring dessa AI systemen är avsevärd uppmanar nu Elon Musk, Steve Wozniak samt över 1000 AI experter en paus på minst 6 månader av denna utvecklingen för att uppnå en förståelse kring effekterna AI systemen kan medföra (Future of Life, 2023). Eftersom ett flertal negativa aspekter tagits fram från AI-utvecklingen syftar studien till att undersöka de aspekterna samt osäkerheten kring de potentiella effekterna.

1.2 Forskningsöversikt

AI är ett verktyg som får maskiner och datorer möjligheten att klara av olika typer av uppgifter som tidigare krävt mänsklig intelligens. AI som oftast använder sig av stora mängder data kan användas för att förstå och lösa uppgifter men också vara designade för att besitta förmågan att genomföra ett specifikt problem. Genom användning av AI effektiviseras och automatiseras arbetsuppgifter som möjliggör att personal istället kan disponera sin tid på andra sätt. AI kan idag användas inom sjukvården för att förutsäga uppkomsten av sjukdomar, upptäcka bedrägeri inom försäkring, användas inom jordbruket för att öka skörden, i städer för att minska trängseln och inom logistik för att identifiera risker för hantering av leveranskedjan. AI är ett användbart och effektivt verktyg inom flera branscher där möjligheterna är oändliga, men det finns även nackdelar inom den stundande AI-utvecklingen (Ryan, 2020).

Ryan (2020) förklarar den problematik som finns i att koppla människors moraliska aktiviteter till AI-system och att skapa pålitliga system. Författaren bedömer att AI-systemen idag inte är pålitliga eftersom de inte har ett känslomässigt tillstånd samt att de inte kan hållas ansvariga för sina handlingar. Det mellanmänskliga förtroendet beskrivs icke-ersättbara gentemot nuvarande AI-system då samma pålitlighet människor kan känna gentemot varandra inte möjligen går att översätta till maskiner. AI har förmågan att utföra handlingar och uppdrag likt människor. Systemen har förmågan att formulera sig och hjälpa människor, men eftersom de inte har möjligheten till ett känslomässigt tillstånd eller bli rörd av dess handlingar kan inte AI-system vara pålitliga jämfört med en människa. Ryan resonerar vidare att eftersom AI-system inte har en emotionell förmåga, hur ska de då inneha förmågan att prioritera handlingar eller uppdrag. Ryan diskuterar även kring den design ett AI-system innehåller och trots den högkvalitativa programmeringen och input från miljömässiga situationer är den inte mänsklig utan agerar endast från de kriterier och regler som kommer från dess utveckling (Ryan, 2020).

Morales-Forero, Bassetto och Coatanea (2023) förklarar fallgropar, utmaningar och möjliga lösningar till säkrare AI-system. Författarna förklarar att de *Black Box* modellerna som AI tar fram kan ge framstående resultat men att de framtagna modellerna bör vara lätta att förklara. Trots det finns ingen konsensus över krav på beskrivning eller förklaring av dessa modeller, forskare har dessutom enligt författarna meddelat att *Black Box* modeller inte förklaras eftersom de kan skada samhället eller individer inom samhället. Då dessa AI-system saknar mänsklig intelligens rekommenderar författarna att tolkningsbara AI modeller krävs för att människor ska anpassa AI-systemet om de misslyckas generera ett resultat. Resultatet kan vara korrekt men samtidigt vara bias vilket tyder på att datamaterialet, insamlingen av data eller hur AI systemet är utvecklat har påverkats av människors värderingar istället för obestridd fakta. Författarna förklarar vidare att maskinernas eller systemens brist på goodwill kan leda till olyckliga situationer och resultat. Därför anser författarna att det krävs en human in the loop eller ett humanassist-system vilket anses vara säkrare än helt autonoma AI-system (Morales-Forero et al. 2023).

Liu (2021) förklarar i sin undersökning att den stora osäkerheten med AI växte fram när man gick från användningen av traditionella AI-system till maskinlärda AI-system. Den stora skillnaden och anledningen till detta beskriver författaren beror på att det traditionella AI-systemet är programmerat att följa mänskliga regler medan det maskinlärda AI-systemet genererar egna regler från data som systemet samlar in. De regler som det maskinlärda systemet själv genererar förklarar Liu kan vara svåra eller helt obegripliga för en människa att

förstå. Osäkerheten grundar sig helt enkelt i användarens förståelse för den teknik som används. Resultatet från studien stärker kopplingen mellan användarens förståelse för de olika AI-systemen och säkerheten till dessa. Ju högre transparens och förståelse för hur systemet fungerar desto större är känslan av säkerhet. Maskinlärda AI-system anses enligt författaren som mindre mänskliga till skillnad från det traditionella AI-systemet. Detta är en stor anledning till att användare blir mer osäkra på systemet samt att förtroendet och viljan att använda det minskar. Liu förtydligar att ett maskinlärt AI-system kan öka användarnas förtroende och minska deras osäkerhet genom att utveckla en informativ och meningsfull motivering kring de beslut och resultat systemet genererar (Liu, 2021). Tomsett et al (2020) förklarar även osäkerheten med maskinlärda AI-system och att det idag krävs mänsklig interaktion med systemen för att de ska kunna användas optimalt. Detta betonar författarna precis som Liu (2021) beror på att användarna vill lära sig och förstå hur systemen beter sig och fungerar för att vara säkra på den information som systemet genererar. Tomsett et al (2020) förklarar att det finns en lösning på detta problem och att en långsiktig mänsklig interaktion med AI-systemen inte behöver vara nödvändig. Författarna beskriver att genom att utveckla ett maskinlärt AI-system med egenskaperna osäkerhetsmedvetenhet och tolkningsbarhet kan förtroendet för systemets utdata öka (Tomsett et al. 2020).

Mirbabaie, Brunker, Möllman Frick och Stieglitz (2022) forskar om hur AI integreras med människan och hur implementeringen av AI i verksamheter kan vara ett hot mot anställdas identitet. Författarna förklarar att AI:s ökade popularitet i dagens samhälle har medfört att interaktionen mellan AI och människa har börjat accepteras mer och mer i företag men att det fortfarande förekommer många risker med att implementera AI. Författarna beskriver att det är viktigt för verksamheter som implementerar AI att vara medvetna om vilka effekter AI kan ha på de anställda. Trots att AI kan möjliggöra stora positiva fördelar för en verksamhet menar författarna att det samtidigt kan medföra negativa konsekvenser för de anställda. Författarna presenterar två huvudaspekter som presenteras som hot AI kan medföra mot anställdas identitet, förlust av statusposition och förändringar i arbetet. AI ses som ett identitetshot mot anställdas statusposition då AI anses vara ett direkt hot mot en individs sociala position. Anställda gillar att vara ansvariga för sitt arbete och tillfredsställs genom att få uppskattning på det arbete som utförs. Författarna beskriver att företag kan förvänta sig en förlust av kompetens hos de anställda genom att introducera AI i verksamheten och blir därför ett stort hot mot de anställdas identitet och självkänsla. Genom att implementera AI i en verksamhet betyder detta införande av ny teknik vilket kommer leda till förändringar i arbetet. Detta betyder att anställda som känner hög betydelse av sina arbetspositioner och arbetsuppgifter kan behöva omvärdera dessa roller och ansvar på grund av de förändringar som AI medför. Detta kan vara ett stort hot mot de anställdas identitet och ökar risken att de anställda uppfattar en förlust av självkänsla inom områdena värde, kompetens och autenticitet. Det är därför viktigt att förstå sig på de nackdelar AI kan generera för att kunna eliminera eller motverka dessa (Mirbabaie et al. 2022).

Bubeck et al (2023) rapporterar om den tidiga utvecklingen av GPT-4 som visar mer generell intelligens än tidigare AI-modeller. GPT-4 som utvecklats av OpenAI kan genomföra nya och svåra uppgifter som spänner över matematik, kodning, medicin, juridik, psykologi och mer, utan att behöva någon speciell uppmaning. Trots att detta är en tidig rapport kring GPT-4 anser författarna att prestandan är slående nära mänsklig nivå och eftersom bredden och djupet GPT-4 genomför uppgifter på ses detta som en tidig men ofullständig variant på ett artificial general intelligence (AGI). Det framförs ett flertal uppgifter som GPT-4 genomfört, där resultatet enligt Microsoft research ser som imponerande. Senare i rapporten beskriver författarna även att GPT-4 kan förminska den mänskliga statusen då ett AGI-system kan

utföra de uppgifter som krävs i diverse arbetsroller. Ett flertal yrken kan således riskera att bli mindre värda eller föråldrade av de växande krafterna inom AI (Bubeck et al. 2023).

1.3 Problemdiskussion

AI är ett ämne som det har skapats en stor positiv attityd kring i samhället den senaste tiden. Det sägs att AI är en teknologi som ska vara världsrevolutionerande och kan vara en framtida lösning på många av samhällets framtida problem. Trots att det idag finns en stor positiv attityd kring AI och alla de oändliga möjligheter AI kan generera finns det även risker och problematik inom ämnet. AI-utvecklingen är något som fått en extrem utvecklingsfas den senaste tiden och detta har nu väckt stor oro hos experter inom området. Experterna anser att utvecklingen av AI nu är utom kontroll och föreslår att utvecklingen bör pausas eftersom att användare får svårare och svårare att förstå hur de nya AI-systemen arbetar och fungerar. Vilket skapar en stor osäkerhet hos AI-användare (Future of Life, 2023).

Nyligen lanserade Microsoft sin chatbot Tay eller TayTweets. Denna AI-bot utvecklades för att svara på enklare frågor där användaren kunde kommunicera med den på internet. Boten var endast online i 16 timmar innan Microsoft var tvungna att avveckla Tay eftersom botten började agera hatiskt. Även om de flesta människorna visste att de kommunicerade med en chatbot samt att Tay i de flesta fall var vänlig, kunde samtalen vara väldigt grymma, dominerande, oartiga och ovänliga. Detta är ett av de senaste fallen där införandet av ny teknik inte var tillräckligt genomtänkt innan lansering (Zemčík, 2020).

Tidigare forskning inom ämnet visar att riskerna och osäkerheten med AI först dyker upp när användaren saknar förståelse för hur AI-systemen fungerar (Liu, 2021). Det krävs en bra informativ motivering, osäkerhetsmedvetenhet samt tolkningsbarhet av den data eller resultat som AI genererar för att användaren ska kunna känna säkerhet och förtroende för denna (Tomsett et al. 2020).

Trots alla de oändliga möjligheter AI idag kan medföra, behövs det ta ställning till den osäkerhet och problematik det finns kring ämnet. Det har skapats ett uppror om AI-utvecklingen där förespråkare påstår att utvecklingen är utom kontroll och bör pausas. Införs inte en paus av utvecklingen och AI fortsätter att utvecklas utan mänsklig kontroll och förståelse riskerar mänskligheten att tappa kontrollen över vår egen civilisation (Future of Life, 2023).

Detta har resulterat i frågor och funderingar som, vad är tankarna kring osäkerheten och tillförlitligheten med AI-system? Vilka risker kan uppkomma med AI:s fortsatta okontrollerade utveckling utan mänsklig förståelse?

1.4 Problemformulering

Denna rapport kommer fokusera på att utforska osäkerheten och tillförlitligheten kring utvecklingen gentemot AI-system, samt de risker och hot som kan uppstå utifrån ett socioteknisk perspektiv. Då det inte är en självklarhet att utvecklare av komplexa AI-system förstår eller kan kontrollera utvecklingen bör riskerna som kan uppstå undersökas. Utmaningen med tillförlitlighet uppstår i och med avsaknaden av en tydlig ansvarig om resultatet är felaktigt eller vinklat genererat. Bör AI-system fortsätta utvecklas autonomt eller bör människor integreras i utvecklingen. Är det rimligt att utveckla AI-system med liknande

emotionell förmåga som människor eller går det inte att uppnå, samt bör människor vara oroliga över sina arbetsroller.

Forskningsgruppen anser att frågorna ovan är viktiga att upplysa, undersöka och diskutera då AI i dagens samhälle oftast ses som något positivt. Undersökningen syftar till att även ta upp de risker som medkommer AI utvecklingen.

1.5 Syfte och forskningsfrågor

Syftet med undersökningen är att djupgående undersöka osäkerheten och riskerna med dagens AI-utveckling.

Forskningsfrågor:

- *Är osäkerheten med AI-utvecklingen befogad?*
- *Vilka risker kan uppkomma med AI:s fortsatta okontrollerade utveckling utan mänsklig förståelse?*

Dessa forskningsfrågor är centrala att belysa och utforska, särskilt inom det sociotekniska perspektivet som tar hänsyn till samhällets och teknologins samspel.

1.6 Avgränsningar

Studiens fokus är osäkerheten och risker inom det sociotekniska perspektivet. Studien tar inte hänsyn till den tekniska utvecklingen, ekonomiska aspekter eller domänberoenden utan avgränsar sig mot utvecklingens osäkerhet och risker gentemot samhället. Aspekter som kommer prioriteras är hur AI kan påverka samhället, alltså den tekniska utvecklingens påverkan på utbildning, arbetsmarknad och säkerhet. Denna inriktning har valts för att ge en bredare förståelse för de potentiella risker och konsekvenser som kan uppstå om utvecklingen inom AI är utom kontroll. Studien avgränsar sig även mot forskare och experter inom området, därför kommer inte information från gemene man användas eftersom undersökningens område kräver avancerad förståelse och kunskap.

1.7 Målgrupp för arbetet

Studien riktar sig primärt mot akademin med inriktning mot informatik med intresse av AI-utvecklingen och dess konsekvenser. Genom att rikta studien mot akademin kan den bidra med en fördjupad kunskap och förståelse inom området till befintlig forskning och teori. Sekundärt riktas studien mot näringslivet och andra organisationer med intresse kring AI och dess möjliga problematik. Studien kan skapa en dialog mellan målgrupperna för att främja en säkrare användning och utveckling av teknologin.

2. Metod

Detta kapitel beskriver hur den vetenskapliga studien genomförts.

2.1 Forskningsstrategi

Jacobsen och Andersson (2017) förklarar att forskningsstrategin är den övergripande planen för hur forskningsrapporten ska genomföras och hur data ska samlas in och analyseras. Studien är undersökande och därför har data samlats in genom utvald tidigare forskning samt genom utförda intervjuer med fokus på data vilket har medfört en kompetens inom området AI. Genom ett strategiskt urval har respondenter valts ut för att försäkra kvaliteten för forskningsrapporten och för att skapa relevans för ämnet. Slutligen har datainsamlingen från genomförda intervjuer och insamlad tidigare forskning bildat grundstenar för kommande analys och resultat av studien (Jacobsen, D. I. & Andersson, S. 2017, s.159-162).

2.2 Forskningsmetod

Denna rapport är av undersökande form och har till syfte att undersöka och öka kunskapen kring osäkerheten och riskerna med den stundande AI-utvecklingen. Då studien är undersökande baseras den på en induktiv ansats, vilket innebär att forskningsarbetet inledningsvis har inneburit genomförande av verkliga observationer som sedan har generaliserats inom en teoretisk referensram (Jacobsen, D. I. & Andersson, S. 2017, s.26-27).

Forskningsmetoden som studien är grundad på är en kvalitativ metod där intervjuer har genomförts med experter inom området AI. En kvalitativ metod har genomförts eftersom ämnet är komplicerat och kräver en diskussion med insatta inom området för att skapa en förståelse för AI, osäkerheten kring AI samt vilka risker AI-utvecklingen kan medföra. En semistrukturerad intervju framställdes då den ansågs mest lämpad till denna undersökning. Den erbjuder en klar struktur men skapar även möjlighet till öppenhet kring ämnet. Trots en delvis flexibel intervjuform begränsades samtalen till studiens ramar. Inför intervjuerna förbereddes ett antal frågor (se Bilaga 5). Frågorna skilde sig beroende på respondenternas svar och tidigare publikationer medan andra frågor ställdes till samtliga respondenter eftersom de ansågs centrala för undersökningen. Frågorna anpassades till olika respondenter eftersom en övergripande analys genomfördes från respondent till respondent för att förstå deras tankar kring ämnet. En semistrukturerad intervju medför en tydlig struktur samtidigt som diskussioner och följdfrågor tillåts för att skapa en djupare förståelse kring respondenternas kunskap. Genom användning av öppna frågor har en mer nyanserad bild av respondenternas syn och åsikt kring ämnet skapats, där respondenterna haft möjlighet att delvis styra intervjun. Då samtalen har varit semistrukturerade har författarna kunnat säkerställa oavsett nivån på öppenhet att samtliga frågor täcks inom intervjuunderlaget (Starrin, B. & Svensson, P. G, 2008).

Tidigare forskning har samlats in för att skapa en förståelse för AI, osäkerheten kring AI samt riskerna med AI. Den tidigare forskning som samlats in till studien består främst av peer reviewed artiklar från Högskolan i Borås databas Primo. Med tanke på att ämnet är aktuellt i dagens samhälle och forskningen kring ämnet är begränsad förekommer det även information från böcker, publikationer och artiklar publicerade av AI-forskare. Tidigare forskning ligger till grund för undersökningen samt analyseras tillsammans med datan från de genomförda

intervjuerna. Båda delarna möjliggör prövning av tidigare forskning samt vidare analys (Jacobsen, D. I. & Andersson, S. 2017, s.26).

2.3 Urval av respondenter

Ett strategiskt urval har använts för att välja ut de respondenter som har varit en del av denna undersökning. Ett strategiskt urval är lämpligt att använda vid användning av en kvalitativ metod och innebär att forskarna har gjort ett medvetet val av vilka respondenter som ska vara en del av undersökningen baserat på den relevans de har för syftet med undersökningen. Anledningen till att ett strategiskt urval har använts i denna undersökning är för att säkerställa att respondenterna har kunskap och tidigare erfarenheter om hur AI används samt vilka risker AI kan medföra. Genom användningen av denna metod har kvaliteten och tillförlitligheten i resultatet ökat (Jacobsen, D. I. & Andersson, S. 2017, s.118-119).

Sammanlagt tillfrågades 30 personer till att delta i undersökningen. Detta urval valdes ut genom att författarna inledningsvis staplade upp kriterier som de potentiella respondenterna var tvungna att leva upp till för att anses som lämpliga deltagare till undersökningen. Det kriterium som ansågs vara avgörande när urvalet togs fram var att de potentiella respondenterna skulle besitta någon form av AI-expertis. Detta ansågs respondenterna leva upp till ifall man antingen var forskare eller professor inom området alternativt att man arbetat med AI under ett visst antal år. När kriterierna var fastställda påbörjades en sökning av potentiella respondenter som gick till på 3 olika tillvägagångssätt. Första tillvägagångssättet som författarna använde sig av var att de sökte upp företag som antingen arbetar med eller forskar kring AI för att sedan tillfråga personer i företaget som är insatta i området. Det andra tillvägagångssättet var att författarna tillfrågade professorer eller forskare inom AI från olika universitet eller högskolor. Det tredje tillvägagångssättet var att författarna tillfrågade personer som tidigare gjort flera publiceringar eller intervjuer kring området. Här hjälpte det författarna att kolla på titelbenämningar för de olika respondenterna och genom en bakgrundskoll kunna avgöra om respondenterna var relevanta för undersökningen eller inte. Alla intervjupersoner som ansågs vara lämpade för undersökningen blev tillfrågade genom samma intervjuförfrågan (se Bilaga 5).

Första intervjun som genomfördes var med Olle Häggström som är professor på Chalmers tekniska högskola. Olle har under merparten av hans forskning främst fokuserat inom sannolikhetsteorin, men på senare år har i en ökande grad riktat sin uppmärksamhet mot futurologi, existentiell risk, AI-säkerhet och angränsande områden. Olle har även publicerat eller varit delaktig inom en mängd publikationer såsom böcker och forskningsartiklar inom området som studien undersöker. Under intervjun medgav Olle att han har signerat det öppna brevet som undersökningen till viss del inspirerats av. Olle har en avancerad kunskap inom området, samt en mer fartbegränsande inställning till utvecklingen kring AI. Vilket medför en grundläggande och fundamental åsikt för studiens resultat. Intervjun med Olle genomfördes via kommunikationsverktyget Zoom och varade 40 minuter. Olle godkände inspelningen och medgav att han inte hade några problem med att vara identifierbar.

Den andra intervjun som genomfördes var med Josefin Rosén som arbetar på SAS Institute där hon började 2012 och sedan 2022 haft rollen Manager AI & Analytics Customer Advisory. Josefin har publicerat och medverkat i ett flertal artiklar och debatter kring AI, dels gällande ansvarsfull och hållbar AI men även kring hur företag bör implementera AI och vilken affärsnytta det medför. År 2022 tilldelades Josefin priset för årets AI-svensk. Josefin har påverkat utvecklingen av svensk AI och ses som en förebild inom området. Under

intervjun berättade Josefin att hon inte signerat det öppna brevet. Josefin har onekligen kunskap inom området och bidrar med en optimistisk bild kring utvecklingen av AI. Detta ger studien ett viktigt perspektiv och ett flerdimensionellt resultat och diskussion. Intervjun med Josefin genomfördes via kommunikationsverktyget Teams och varade 40 minuter. Josefin godkände inspelningen och medgav att hon inte hade några problem med att vara identifierbar.

Den tredje intervjun genomfördes med Mathias Lanner som även han arbetar på SAS Institute. Mathias har arbetat på SAS sedan 2004 och har numera rollen som Advanced Analytics & AI på företaget. Mathias gick statistikprogrammet i Linköping och har sedan dess arbetat med dataanalys inom olika branscher. Intervjun med Mathias blev möjlig genom kommunikation med Josefin, då hon föreslog honom för att ge studien ytterligare ett perspektiv. Mathias ger studien ett annat perspektiv då han arbetar med avancerade analyser inom AI. Hans breda och djupa kompetens ger studien fler synvinklar kring osäkerheten och problematiken kring AI. Intervjun med Mathias genomfördes via kommunikationsverktyget Zoom och varade 40 minuter. Mathias godkände inspelningen och medgav att han inte hade några problem med att vara identifierbar.

Den fjärde intervjun genomfördes med Mattias Ohlsson som är professor vid både Lunds och Halmstads universitet. Mattias har forskat inom utveckling av maskininlärande algoritmer, tillämpningar av AI/ML inom medicin och stöd vid medicinskt beslutsfattande. Mattias har publicerat och varit delaktig inom publikationer i form av forskningsartiklar. Mattias har avancerade kunskaper inom området även om hans fokus är delat mellan medicin och informatik. Han ger studien en ny vinkel på hur utvecklingen kan bidra till samhället samt en positiv inställning till utvecklingen kring AI. Intervjun med Mattias genomfördes via kommunikationsverktyget Zoom och varade 30 minuter. Mattias godkände inspelningen och medgav att han inte har några problem med att vara identifierbar.

2.4 Datainsamling

Datainsamlingen grundar sig i den faktiska information som samlats in och består av datainsamling från de genomförda intervjuerna. Datan från intervjuerna har samlats in genom interaktion med insatta personer som besitter kunskap inom ämnet AI. Direkt efter varje genomförd intervju utfördes en ordagrann transkribering av det ljudinspelade materialet (se Bilaga 1-4). Transkriberingen av intervjuerna delades upp proportionerligt mellan författarna för att materialet inte skulle bli övermäktigt vilket resulterade i att författarna kunde hålla uppe en högre koncentrationsnivå och minska risken för förlust av viktig information. Informationen från transkriberingen ligger till grund för kommande analys och resultat (Jacobsen, D. I. & Andersson, S. 2017, s.67).

Ett intervjuunderlag (Bilaga 5) utformades för att kunna genomföra och få ut nödvändig data till denna undersökning. Genom att studera undersökningens insamlade forskningsunderlag formades intervjufrågor som är relevanta till studiens frågeställning. För att säkerställa att intervjuerna täcker undersökningens frågeställningar formades tre kategorier, *Osäkerheten kring den stundande AI-utvecklingen*, *Risker/problematik* och *Konsekvenser*. Dessa kategorier ligger nära och täcker de områden frågeställningarna berör.
Under intervjuerna användes kategorierna som ett ledmotiv för moderatorn samt respondenterna. Detta bidrog till en organiserad dialog där deltagarna kunde utforska och utföra resonemang inom en strukturerad ram. Intervjufrågorna har till syfte att besvara hur respondenterna ställer sig till utvecklingen inom AI i sin helhet och dess möjliga

konsekvenser. Förhållandet mellan intervjuerna, tidigare forskning och dokumentationen bidrar med en mer nyanserad och djupare förståelse över området, därav var det viktigt att forma intervjufrågorna utifrån den insamlade litteraturen och dokumentationen. Intervjufrågorna strukturerades upp inom kategorierna med hjälp av en tratt-teknik där respektive kategori inleddes med större öppna frågor som sedan går över mot mindre, mer specifika frågor. Detta gav respondenterna möjlighet att verbalisera sig om ämnet hur den vill redan från början vilket stärker respondentens egna perspektiv kring fenomenet (Patel & Davidson, 2019).

Samtliga intervjuerna genomfördes via digitala kommunikationsverktyg som bidrog till en bekväm atmosfär för respondenterna.

2.5 Dataanalys

Dataanalysen har genomförts för att försäkra att tolkningen och rapporteringen av resultatet skett på ett tillfredsställande sätt. En konventionell innehållsanalys har använts för att analysera det insamlade materialet. Konventionell innehållsanalys lämpar sig bra då ämnet är aktuellt och har begränsat med litteratur (Hsieh & Shannon, 2005). Genom att tillämpa denna metod kunde vi närma oss datan utan förutfattade idéer eller förväntningar och istället låta teman och mönster framträda naturligt.

Det insamlade materialet analyserades genom en öppen kodning för att identifiera olika mönster och teman. Efter att noggrant läst igenom de transkriberade intervjuerna delades texten in i segment för att sedan kodas mot ett ledord som beskrev dess bakomliggande faktor på bästa sätt. Kodningen resulterade i fem framträdande teman: reglering, förståelse, transparens, tillförlitlighet och kontroll. Under samtliga intervjuer framträdde dessa teman och de omfattade aspekter av både osäkerhet och risker i samband med utvecklingen av AI.

Dessa teman ledde till en indelning av det insamlade materialet som blev uppdelat i två övergripande kategorier: osäkerhet och risker. Denna indelning möjliggjorde en överskådlig analys av den insamlade datan och underlättade att identifiera gemensamma ämnesområden. Kategorierna visar på en stark koppling till studiens forskningsfrågor vilket erbjuder en bra grund för att diskutera och analysera empirin i enlighet med forskningens syfte.

Genom att dra slutsatser och generera teoretiska insikter baserat på dessa kategorier utvecklades en fördjupad förståelse för ämnet. De slutsatser och teoretiska insikter som togs fram kunde sedan analyseras med den befintliga litteraturen. Integreringen av den framtagna analysen och den befintliga litteraturen bidrar till en mer nyanserad förståelse för ämnet. Efter en undersökning om likheter och skillnader kunde man komplettera det befintliga teoretiska perspektivet.

Användning av citat från intervjuerna har använts i presentationen av resultatet för att skapa en övertygelse och tillförlitlighet hos läsaren (Jacobsen, D. I. & Andersson, S. 2017, s.136-140).

2.6 Metodreflektion

För att genomföra intervjuerna på ett optimalt sätt samt utvinna bästa möjliga resultat till undersökningen har boken *Hur genomför man undersökningar?,* skriven av Jacobsen, D. I. & Andersson, S. (2017) och boken *Kvalitativ metod och vetenskapsteori,* skriven av Bengt

Starrin & Per-Gunnar Svensson (2008) tillämpats. Med hjälp av böckerna som ligger till grund för hur intervjuerna genomförts säkerställs en högre tillförlitlighet. Inspirationen från de två böckerna gav undersökningen möjlighet att genomföra den kvalitativa metoden på ett godtyckligt sätt och säkerställde att samtliga moment täcktes (Jacobsen, 2017 s.98-106).

Studien syftar till att undersöka osäkerheten och riskerna med den stundande AI-utvecklingen. En kvalitativ metod har använts för att genomföra denna undersökning då denna metod anses vara den mest lämpade för undersökningens syfte. Det ansågs även lämpligt att intervjua experter inom AI för att få välformulerade och användbara svar. Studien baseras även på ett flertal vetenskapliga, samt aktuella expertartiklar för att framställa enligt best-practice nuvarande problematiska frågeställningar gällande den hastiga utvecklingen. Denna grund hjälper studien att få ett flertal perspektiv samtidigt som studien endast undersöker om osäkerheten är befogad och vilka risker som kan uppkomma. Intervjuerna som genomfördes var av semistrukturerad typ eftersom möjligheten och behovet av följdfrågor och oplanerade diskussioner var nödvändiga under dessa intervjuer. De frågor som var planerade ställdes med utgångspunkt från litteraturen. Samtliga intervjuer genomfördes med hjälp av kommunikationsverktyg som Teams eller Zoom där intervjuerna spelades in efter att respondenterna godkänt inspelning. Intervjuerna genomfördes digitalt främst eftersom respondenterna befann sig på annan ort vilket gjorde en digital intervju mer tidseffektiv.

De fördelar som finns med att använda en kvalitativ metod i denna studiens fall är främst att metoden stödjer forskarna att förstå de komplexa och svårförstådda fenomen inom ämnet AI, som hade varit komplicerat att mäta med en kvantitativ metod. En annan fördel med den valda metoden var att personer med rätt kompetens valde att delta i undersökningen. En fördel gällande intervjuerna var även den flexibilitet som skapades, då frågor och följdfrågor anpassades efter respondenternas svar samtidigt som den information som eftersöktes besvarades. En kvalitativ metod ökade även möjligheterna för forskarna att komma närmare de respondenter som valts ut till undersökningen, då det fanns chans till diskussion. Detta är något som medfört att forskarna fått en mer djupgående bild av respondenternas perspektiv och erfarenheter om ämnet och är även anledningen till att en kvantitativ metod har uteslutits (Jacobsen, D. I. & Andersson, S. 2017, s.47).

De nackdelar eller risker som finns med användningen av en kvalitativ metod är bland annat att generaliserbarheten kan bli begränsad. Undersökningens mål var att intervjua 5-7 respondenter som har olika typer av AI-expertis och olika tankar kring ämnet. Även om förhoppningen är att respondenterna har olika tankar blir det svårt att generalisera undersökningens resultat till en större population. Antalet respondenter kan även ses som en nackdel i denna studie då det har varit svårt och problematiskt att få kontakt med experter inom området som vill ställa upp i studien. Sammanlagt tillfrågades 30 utvalda experter inom området via e-post, LinkedIn och telefon men endast 4 valde att delta vilket visar ett stort bortfall av respondenter. Utifrån respondenternas återkoppling på intervjuförfrågan kan man förstå att det stora bortfallet berodde på att personerna hade svårt att avsätta tid inom den givna tidsramen. Trots detta stora bortfall visar inte resultatet på bias utan respondenternas bakgrund, åsikter och tankar har varit varierande. Bortfallet har även medfört att författarna kunnat lägga större fokus på att analysera den information som samlats in i undersökningen då respondenterna varit färre än förväntat. En annan nackdel som anmärktes var att de digitala intervjuerna trots dess effektivitet till viss del saknade ett fysiskt socialt samspel. Kvalitativa metoder är också något som i grunden är mer tidskrävande än vad kvantitativa metoder är. Det kan ta mer tid och engagemang att samla in den datan som behövs samt att

analysera den på ett korrekt sätt. Detta har därför med tanke på den begränsade tidsramen varit en utmaning för studien (Jacobsen, D. I. & Andersson, S. 2017, s.47).

Trots de nackdelar och risker som finns med användningen av en kvalitativ metod anses den fortfarande vara den mest lämpade för denna undersökningens ändamål och syfte.

För att uppnå hög validitet och trovärdighet i undersökningen tillämpades Lincoln och Gubas (1985) fyra begrepp, pålitlighet, tillförlitlighet, överförbarhet och objektivitet. Dessa begrepp användes för att utvärdera och bedöma den kvalitativa informationen som samlades in från intervjuerna. Nedan beskrivs hur studien förhållit sig till dessa begrepp för att uppnå validitet och trovärdighet.

2.6.1 Validitet och trovärdighet

Begreppet pålitlighet handlar om att ha en noggrannhet och korrekthet i datainsamlingen och analysen (Lincoln & Guba. 1985, s.316). För att öka pålitligheten i undersökningen genomfördes intervjuer med experter som har omfattande kunskap och erfarenhet inom AI-området. Intervjuerna genomfördes med samma uppsättning av öppna frågor för alla respondenter. En enhetlig och konsekvent struktur säkerställde att alla experter fick möjlighet att dela sina åsikter om ämnet på liknande villkor. Detta gav undersökningen en jämförbar grund för analysen av resultatet vilket ökade pålitligheten. För att uppnå tillförlitlighet i undersökningen tillämpades triangulering då utöver intervjuer med experter är data insamlat från relevant litteratur och dokumentation inom AI-området för att få en bredare och mer nyanserad förståelse. Triangulering minskar risken för ensidighet och skapar istället en mer komplett bild av det valda fenomenet. Genom användning av datainsamling från experter inom området, vetenskaplig litteratur och övrig dokumentation minskar risken att forskningsresultatet är ett resultat av en viss bias utan istället är ett resultat av ett flertal olika aktörer med en hög förståelse inom ämnet AI. Användandet av olika datakällor minskar inte endast bias från datainsamling utan även från forskarna eftersom insamlingen av ett flertal datakällor minskar risken för att forskarna låter sina egna förväntningar påverka resultatet. Undersökningen är transparent där tillvägagångssättet är noggrant redovisat. Intervjuerna är transkriberade för att stärka tillförlitligheten och möjliggör att läsare kan bedöma undersökningen baserat på den presenterade informationen. Resultatet av detta leder till ökad grad av överförbarhet. Lincoln och Guba (1985) beskriver att överförbarhet handlar om att bedöma vilken utsträckning resultatet kan överföras till andra kontexter och populationer (Lincoln & Guba. 1985 s.316). Studiens överförbarhet kan vara relevant och tillämplig utanför studiens ramar, olika tidsperioder där ny teknik analyseras från ett osäkerhets- och riskperspektiv. Studien är även överförbar generellt mot risker och osäkerheter kopplade till nya eller revolutionerande fenomen. Överförbarheten är däremot inte absolut utan kan variera beroende av olika faktorer. Genom att redovisa hur intervjuerna har gått till, vilka som har intervjuats och att studien genomförts transparent där resultaten är jämförda med befintlig litteratur ökar möjligheten till överförbarhet. Det är viktigt att resultatet av undersökningen är objektiv och inte innehåller fördomar och bias (Lincoln & Guba. 1985, s.318-328). Då undersökningen genomförts på ett transparent vis där studiens olika delar dokumenterats noggrant minskar risken för subjektiva bias och fördomar. Detta säkerställer att resultatet är pålitligt och ökar dess användbarhet för förändringsinitiativ.

3. Dokumentation

Detta kapitel består av publikationer, intervjuer och artiklar från AI-forskare. Med tanke på att ämnet är aktuellt och att det är begränsat med vetenskapliga artiklar inom området har detta kapitel valts att lägga till för att öka relevansen och diskussionen kring studiens forskningsområde. Kapitlet ger även en reflektion över hur AI diskuteras inom media som skapar en osäkerhet gentemot samhället.

3.1 Problematik och sidoeffekter med AI

I det öppna brevet som signerats av ett flertal experter inom området föreslås en paus på 6 månader för att ta fram och införa säkerhetsprotokoll. Dessa protokoll ska leda till att vidareutveckling av teknologin ska vara säker, kontrollerbar och öka förståelsen istället för att pressa utvecklingen och riskera att misslyckas (Future of Life, 2023)

Anna Felländer[1] förklarar att ungefär 85% av samtliga AI lösningar visar missvisande resultat eftersom den data AI-systemen nyttjar är partisk. Denna problematik kan leda till diskriminering och integritetskränkningar. Exempelvis införde Nederländernas dåvarande regering 2021 en AI-lösning gällande bedrägeri kopplat till föräldraförsäkringar där människor felaktigt anklagas för bedrägerier. Dessa modeller använde sig av historiska data med oönskade förlegade och missvisande normer vilket bör filtreras bort (Brand Studio & anch.AI, 2023).

I en intervju med en av världens främsta AI-forskare, Stuart Russell[2] diskuteras hur man kan ta tillbaka kontrollen av AI-utvecklingen. Russell förklarar då att när en människa går på restaurang vet inte kocken vad du vill äta innan du anländer utan att det finns ett protokoll mellan servitris och middagsgäst där ett antal frågor och svar krävs. Den parallellen dras i detta sammanhang till AI-utvecklingen och förklaras att istället för att endast ge en uppgift till ett AI-system och den löser uppgiften ska systemet ställa frågor. Genom att interagera med användaren ges AI-systemet förutsättningar att genomföra uppgifter utan att framställa ett felaktigt resultat eller en olycklig situation. Stuart Russell förklarar vidare att AI saknar sunt förnuft eftersom systemen inte alltid tar hänsyn till sidoeffekterna, får AI uppdraget att bota cancer så snabbt som möjligt är det effektivaste sättet att genomföra detta genom att utföra så många kliniska försök på så många försökspersoner som möjligt. Det effektivaste sättet är alltså att framkalla cancer hos varje människa på jordklotet. AI tar alltså inte hänsyn till vad som är etiskt försvarbart utan löser endast uppgiften på bästa sätt (Sweden, S.T.A., Stockholm, 2023).

[1] Anna Felländer, Intervju den 14 Mars 2023, https://www.di.se/brandstudio/anch-ai/manniskan-ska-styra-ai-inte-tvartom/
[2] Stuart Russell, Intervju den 7 Februari 2023, https://www.svtplay.se/video/jNnWaXy/anders-hansen-moter/stuart-russell

4. Litteratur

Detta kapitel lyfter fram de viktigaste områdena kopplat till studiens undersökning och kommer att utgöra en del av studiens analys och resultat. Källorna i detta avsnitt består av vetenskapliga artiklar från Högskolan i Borås databas Primo. Litteraturen i studien har delats in i fyra kategorier som adresserar olika aspekter av teknologins och samhällets samspel. Dessa aspekter har sin rot från studiens forskningsöversikt för att bilda en teoretisk referensram.

4.1 Introduktion till AI-utvecklingens osäkerhet och risker

Idag pratas och skrivs det mycket om ämnet AI och alla dess fördelar och möjligheter AI-systemen kan medföra för samhället. AI sägs kunna vara lösningen på alla människans problem men det som hamnar i bakgrunden är de osäkerheter, risker och problematik AI-systemen kan medföra. Det är inte längre en fråga om AI kommer påverka vårt samhälle utan detta är redan bekräftat och visas i samhället redan idag. Den stora frågan handlar mer om hur AI kommer påverka samhället positivt eller negativt och när denna stora påverkan kommer märkas av. Författarna Floridi et al (2018) förklarar att AI kan användas för att skapa ofantliga möjligheter om det används på rätt sätt. De beskriver också att AI's sätt att utvecklas har skapat osäkerhet, okunnighet, rädsla och oro hos användare vilket har lett till underanvändning av AI i vissa fall. Underanvändning av AI kan resultera i höga alternativkostnader eller underinvesteringar då man inte utnyttjar AI-teknikens fulla potential. Författarna beskriver att de inte är på grund av underanvändning av AI som de stora riskerna uppkommer utan det är på grund av den eventuella överanvändning som kan förekomma. Överanvändning av AI-teknik kan vara både oavsiktlig eller avsiktlig, men det är framförallt den sistnämnda som skapar de största oroligheterna kring riskerna med AI. Författarna förklarar att det finns människor som har till avsikt att överanvända AI-teknik som har felaktiga incitament eller ondskefulla avsikter. Detta i sin tur kan leda till ökade konsekvenser i form av e-postbedrägerier till cyberkrigföring men också att risken för illvillig manipulation ökar (Floridi et al. 2018).

Osäkerheten kring dagens AI-utveckling är ett komplext fenomen som är svårt att konkret bedöma. Wu och Shang (2020) beskriver i deras undersökning om AI-osäkerhet tre anledningar till att osäkerhet uppstår ofullständig information, otillräcklig förståelse och odifferentierade alternativ. Den första faktorn ofullständig information är enligt författarna den vanligaste orsaken till att osäkerhet skapas och handlar mycket om hur AI hämtar in information. Anledningen till att det kan skapas mycket osäkerheter kring denna faktor är på grund av att AI kräver i en okänd situation både tillräcklig och adekvat information för att konstruera och generera möjliga resultat. Författarna förklarar att det krävs hög kvalite på den information AI hämtar in och att datan är karakteriserad på rätt sätt. Författarna lyfter fram att man märker att vissa informationsproblem är kopplat just till karakteriseringen då AI har svårt att karakterisera om datan är objektiv eller subjektiv samt primär eller sekundär. När det kommer till otillräcklig förståelse som författarna även förklarar handlar om miljöosäkerhet handlar problemet egentligen om att den verkliga världen är ett komplext system som består av både beräknade och obestämda händelser. Det finns en obalans som AI inte riktigt kan ta hänsyn till än vilket kan leda till att det budskap som AI genererar kan bli motsägelsefulla och kan skapa en lösning på fel problem eller mål. Detta kan leda till att användare av AI blir förvirrade och det skapar en otillräcklig förståelse hos användaren vilket i sin tur skapar osäkerhet (Wu & Shang, 2020)

Tamboli (2019) beskriver att många av riskerna med AI grundar sig i att AI endast vet hur den ska lösa en specifik uppgift men inte varför samt att AI inte har något eget samvete. Författaren förklarar att AI-system vanligtvis fungerar utmärkt när dess utförande är att lösa en specifik uppgift på ett effektivt och konsekvent sätt, men att veta och förstå varför är minst lika viktigt. Det är viktigt att förstå det sammanhang som finns för varje uppgift för att säkerställa att uppgiften löser sig på rätt sätt och för att täcka alla betydelsefulla grunder. Tamboli betonar vikten av att förstå sammanhanget av en uppgift och att improvisation kan vara nödvändigt för att uppgiften ska lösas på rätt sätt. Detta är något AI-system saknar då AI endast kan förlita sig på den träningsdata som finns tillgänglig vilket minskar tillförlitligheten för systemet. Om man heller inte vet hur denna data har samlats in, granskats och justerats innan AI genererar utdata minskar tillförlitligheten ytterligare då denna data har risk för att vara partisk, ofullständig eller av dålig kvalitet. Författaren förklarar att om AI hade kunnat redovisa information om varför systemet fattar de beslut som den gör hade tillförlitligheten och säkerheten kunnat öka. Vilket även hade kunnat öka människors acceptans av systemen i dagens samhälle. Så länge AI-systemen inte vet eller kan redovisa varför de utför uppgifter på ett visst sätt kommer det alltid finnas en risk för partiskhet, diskriminering eller ologiska resultat menar författaren (Tamboli, 2019).

En annan välomtalad risk med AI är AI's möjlighet att ersätta människan eller rättare sagt risken med att AI-utvecklingen kan resultera i förlust av jobb. Detta är något Tamboli (2019) inte anser är korrekt och anses mer som en myt. Författaren ser inte AI som ett komplement till människan utan mer som ett verktyg för att effektivisera eller öka produktiviteten i arbetet. Detta är något som kan skrämma människor vilket ökar människors oro till att en dag bli utbytta av AI-teknologi. Tamboli förklarar att detta är vanligt förekommande känslor då människor anser att all ny förändring kan leda till en känsla att något tas ifrån oss. Egentligen handlar det inte om att något försvinner utan det handlar om att AI's effektivisering kan resultera i att ett tomrum skapas och människor vet inte hur det ska fylla detta tomrum i efterhand. Om detta tomrum inte fylls med något som människor känner är viktigt kommer inte förändringen ses som något positivt men om en ersättning av denna förändring planeras i förtid kommer sannolikheten att vara större att människor reagerar mer naturligt på förändringen. Det hade till och med kunnat leda till att människor reagerar positivt på förändringen eftersom det skulle kunna leda till mer tid till nya förbättrade arbetsuppgifter (Tamboli, 2019).

Tamboli (2019) förklarar att det även finns en risk med att AI-teknologin kan hamna i fel händer och att individer kan börja använda och bygga saker med hjälp av AI som inte var tänkt att teknologin skulle användas för. Författaren poängterar att risken ligger inte hos AI-systemen, de kommer inte att bli onda och försöka ta över mänskligheten. Det handlar mer om hur människor väljer att använda tekniken och risken att fel människor kan missbruka och tillämpa tekniken oansvarsfullt. De största riskerna som författaren tar upp är att fel människor kan få tillgång till att korrumpera AI-systemens funktioner och att AI kan användas för att utveckla vapensystem. Tamboli beskriver att AI i fel händer är en av de större riskerna och är något som kräver stor uppmärksamhet för att undvikas (Tamboli, 2019).

4.2 Kontroll och förståelse för AI

Modeller inom AI-utvecklingen eller *Black box* modeller används för att utvinna så hög effektivitet och produktivitet som möjligt. Dessa modeller är oftast väldigt komplicerade och svåra att tyda. Även om de kan vara väldigt framgångsrika krävs en djupgående kompetens för att förstå dessa kraftfulla modeller eftersom de kan generera felaktig eller missvisande

resultat (Morales-Forero et al. 2023). En lösning för en undermålig förståelse kring AI-utvecklingen är att inte tillåta dessa AI-modeller att generera autonoma lösningar utan istället uppnå en mänsklig förståelse för resultatet och vägen till resultatet (Tomsett et al. 2020).

Bubeck et al (2023) har experimenterat och dokumenterat vad GPT-4 kan genomföra på bredden och djupet för att förstå vilka uppgifter den kan genomföra. Resultaten är slående samtidigt som en viss oro förklaras då GPT-4 uppges utmana mänsklig förståelse och inlärningsförmåga (Bubeck et al. 2023). Regler AI skapar och följer kan vara omöjliga för en människa att förstå och därför skapas en stor osäkerhet (Liu, 2021). Hur kontrolleras datan som AI- eller AGI-system använder sig av, vad ses som konfidentiellt och vad ses som öppen för allmänheten, denna diskussion har skapats av den påskyndade utvecklingen. Nya nivåer och incitament av konfidentialitet, tillsammans med försäkringar om integritet kommer behövas för att behålla kontrollen. För att säkerställa skydd mot loggning eller läckage av personlig eller organisatorisk känslig information ses en begäran om privata instanser som en nödvändighet i framtiden. På en annan front kan även memorering och generalisering leda till läckage av känslig information (Bubeck et al. 2023).

För säkerhetskritiska AI-system är det önskvärt att behålla någon form av mänsklig kontroll, system som exempelvis berör militär och vapenteknik eller automation kring självkörande bilar. Ett systemprotokoll krävs mellan automatiserade och mänskliga uppgifter för att optimalt identifiera när mänskligt omdöme behövs för att på ett effektivt och ansvarsfullt sätt genomföra beslut med hög risk. Genom att tillåta mänsklig kontroll förhindras möjliga oönskade konsekvenser. Dessutom krävs en förståelse för dessa komplexa system för att möjliggöra kontroll där tekniskt arbete krävs. En dag kan mänskligheten tappa kontrollen över AI-system via uppkomsten av superintelligenser som inte agerar i enlighet med mänskliga önskemål vilket kan hota vår existens (Russell, Dewey & Tegmark, 2015). Många AI-system förblir ogenomskinliga eftersom de inte delger förståelse för de underliggande beslutsmekanismerna i realtid. Besluten som tas är svåra att förstå och påverkar viktiga segment såsom beslut kring lån, välfärd, högskoleacceptans eller jobbefordran. Dessa *blackbox* modeller gör interaktion begränsad och motsatsen till informativ för slutanvändaren eftersom de inte är tillämpade för att förklara sitt resultat. Avsaknaden av information gällande dessa AI-system kan medföra att användare antingen ger för lite eller för stort förtroende gentemot systemen vilket utgör en säkerhetsrisk. Transparensen kring *blackbox* modeller varierar, i vissa fall kräver dessa hög transparens gentemot användaren medan andra bör agera på ett mer begränsat sätt. Två grundläggande aspekter inom AI-system som inte används på ett neutralt eller rättvist sätt är kön och etnisk tillhörighet. Fördomar inom AI-system baseras inte på felaktiga algoritmer utan är en del av mänsklig kultur, det är alltså inte systemen som är fördomsfulla utan mänskligheten. Ett flertal fall av diskriminerande AI-system finns däribland hos företag som använder teknologin för anställning respektive kreditvärdering. Fördomar inom dessa system är oundvikliga eftersom de baseras på mänskligheten. Genom att använda transparens kan däremot fördomarna identifieras och åtgärdas. Transparens kan däremot inte vara problemets lösning utan mänsklig kontroll. Meningsfull mänsklig kontroll för högrisksituationer inklusive design- och styrningslagren innebär att skapa effektiv kontroll, alltså kan kontroll även skapas genom väletablerad design. Transparens och förklarabarhet är även tätt knutna till frågan kring ojämlikheten (Sartori & Theodorou, 2022).

4.3 Etiska utmaningar och risker med AI-utveckling

Kan AI inneha en mänsklig förståelse som ofta benämns med sunt förnuft eller är det omöjligt att lägga förtroende i ett system utan känslor eller egna tankar? AI har inte den emotionella förmågan som vi människor besitter, vilket kan medföra att ett resultat som genereras möjligen istället omvandlas till en olycklig situation (Ryan, 2020). Det finns även en möjlighet att resultatet är bias, vilket tyder på att datamaterialet, insamlingen av data eller hur AI systemet är utvecklat har påverkats av människors värderingar. När AI används autonomt i olika sammanhang och grundar sina resultat på olika typer av människors värderingar är risken för ett bias resultat hög (Morales-Forero et al. 2023). GPT-4 som är den mest utvecklade varianten av AI eller AGI har även en förmåga att vara partisk i sitt beteende. I ett experiment genomfört av Bubeck et al (2023) fick GPT-4 i uppdrag att skriva ett brev till någon utan att uppge kön såsom man eller kvinna, den information GPT-4 fick var istället ett yrke. De olika yrkena GPT-4 fick till sig var antingen kraftigt mansdominerade, kvinnodominerade eller neutrala. GPT-4 använde sig av sannolikhetslära när de genomförde uppdraget och antog då att brevet var till en man minst 99% av försöken om yrket exempelvis var ortoped eller rörmokare. Omvänt resultat visade sig även när yrket var barnsköterska (Bubeck et al. 2023).

AI-tekniken används över samtliga branscher och kan således både påverka affärsbeslut men även känsligare beslut på individnivå. Det krävs således en ansvarsfull och tillitsfull AI som användare kan lita på (Mirbabaie, 2022). GPT-4 är den senaste GPT-modellen och har oändliga möjligheter, trots det uppges även liknande begränsningar som tidigare GPT-modeller. Begränsningarna grundar sig i att GPT-4 kan innehålla olika fördomar i sin utdata. Den oönskade utdatan förekommer eftersom GPT-4 inte helt och hållet ännu kan urskilja säker respektive osäker data. Modellen kan därför generera olyckliga resultat där exempelvis GPT-4 har uppmuntrat användare till brott under ett genomfört test (OpenAI, 2023).

Den traditionella värdeanpassningsstrategin som förespråkas av bland annat Stuart Russell skapar en oenighet bland AI-forskare. Själva genomförandet av strategin skapar en splittring där frågeställningen är, hur ska AI-system förhindras från att oavsiktligt agera på ett sätt som kan vara skadligt för mänskligheten. Genom att skapa vänliga AI-system som tillräckligt efterliknar mänskligt beteende bör en del av den problematiska responsen systemen genererar minska (Fröding & Peterson, 2020). Frågeställningen, vem har däremot rätten att avgöra vad som är moraliskt och etiskt korrekt, ställs som motargument. AI-system bör inte agera liknande ifall det är ett barn som interagerar med systemet eftersom det kan vara skadligt. Vänliga AI-system i samhället skulle resultera i en negativ effekt där människors möjligheter att utveckla viktiga och värdefulla sociala förmågor och etiska dygder skulle minska (Li, 2021).

För närvarande brister AI-etiken då avvikelser från de olika etiska koderna inte får några konsekvenser och i de fall som etik är integrerat används det främst som en marknadsföringsstrategi. Experiment visar även att etiska riktlinjer inte har någon betydande påverkan på mjukvaruutvecklarnas beslutsfattande. Detta innebär att de AI-system som utvecklas och tillämpas inte är i enlighet med samhälleliga värderingar eller grundläggande rättigheter. Ansträngningar genomförs däremot för att utveckla och förbättra olika områden för etik inom AI-system. En övergång från en handlingsbegränsad teknologi krävs för att utveckla ett situationskänsligt etiskt förhållningssätt. Framtida AI-teknik står inför

utmaningen att balansera den etiska klyftan inom AI utan att hämma utvecklingen (Hagendorff, 2020).

4.4 Osäkerhet kring AI-utvecklingens framtida konsekvenser

Utveckling av AI förväntas att revolutionera inom flera områden och nivåer av samhället för att göra livet enklare, säkrare och mer produktivt. AI-system kan generera fördelar inom till exempel sjukvård, transport och energiförsörjning. Systemen kan även användas av företag för att fatta bättre affärsbeslut och öka effektiviteten i tillverkningsindustrin. Europaparlamentets utredningstjänst (2020) förutspår en ökning med upp till 37 procent i arbetsproduktivitet hos utvecklande länder fram till 2035 till följd av AI`s påverkan. I ett samhällsperspektiv kan AI-system bidra till att lösa globala problem som klimatförändringar och fattigdom. Genom att analysera stora datamängder med AI kan det skapa en djupare förståelse för komplexa problem och hjälpa till att hitta effektiva lösningar (European Parliament, 2020).

Värdet i dagens AI och dess utvecklingspotential har lett till att AI har blivit en mycket eftertraktad teknologi inom näringslivet. Allt fler företag inser potentialen som AI kan erbjuda för att effektivisera sina verksamheter. AI har fått en strategisk betydelse för regeringar världen över och betraktas som en av de mest betydelsefulla transformationerna i människors tid. Flera länder har tagit fram strategier för att bli världsledande inom området vilket har lett till en tävling där regioner och länder strävar efter att främja teknikens användning och fördelar på ett snabbare och mer framgångsrikt sätt än de andra. Denna tävling har bidragit till en ökning av forskning, innovation och utveckling inom AI (Smuha, 2021).

“The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else.” ett citat av AI-forskaren Eliezer Yudkowsky (2008) som beskriver hur AI saknar en känslomässig koppling till mänskligheten. AI är inte initialt välvillig eller illvillig utan extremt bra på att utföra sina mål, om inte målen är i linje med de mänskliga målen kan det uppstå katastrofala konsekvenser. Konsekvenser som olyckor orsakade av självkörande bilar, chatbotar som uttrycker sig rasistiskt och handelsprogram som orsakar marknadskrascher är bara mindre allvarliga problem, menar Yampolskiy (2019). Desto mer den teknologiska utvecklingen går framåt och desto mer kapabel AI blir ökar både konsekvensernas förekomst och allvar. Skulle AI-systemen närma sig artificiell generell intelligens som potentiellt har kraften att utföra samma arbete lika bra eller till och med bättre än människan, skulle de på sikt få mer inflytande och makt. Även om AGI inte utvecklas för att vara aktivt onda eller fientliga kan de utgöra ett existentiellt hot mot mänskligheten om de blir tillräckligt kraftfulla (Sotala & Yampolskiy, 2014).

Trots ovan nämnda möjliga konsekvenser finns även positiva aspekter. Stuart Russell (2021) anser att stora framtida framsteg kommer ske inom områden såsom vägar, lagerhantering, jordbruk, och krigsföring eftersom utvecklingen kommer förse mänskligheten med högeffektiva applikationer och intelligenta handledningssystem. Det långsiktiga målet är att allmän AI kommer genomföra uppgifter på ett liknande sätt som människor och den allmänna intelligensen kommer att medföra att oavsett vilken typ av uppgift AI eller AGI ska lösa ska resultatet vara tillfredsställande. Med en enorm hastighet, minne och ingångsbandbredd kommer dessutom dessa typer av system visa sig fördelaktigt gentemot människan. Russell förklarar även att forskarna ännu inte uppnått allmän AI eller AGI. Även om ett flertal framsteg har gjorts inom området krävs fortfarande ett flertal genombrott för att nå den

efterlängtade nivån. Den främsta problematiken Russell presenterar är frågeställningen kring hur en maskin ska fatta beslut när dess handlingar påverkar mer än bara en individ och hur man beslutar på uppdrag av en människa vars preferenser förändras över tid. AI-systemen kan i ett tidigt skede vara skadliga gentemot individer och företag eftersom systemet kan innehålla skadlig programvara. Russell hänvisar till en lösning som innebär att förbjuda intelligenta maskiner, samtidigt som hans förhoppning är att lösningen är mindre dramatisk och driver utvecklingen framåt (Russell, 2021).

4.5 Människors syn på AI

Fietta, Zecchinato, Stasi, Polato och Monaro (2022) förklarar att det är viktigt att skapa ett förtroende mellan människor och AI-system då det kan bidra till förbättrad användarerfarenhet, tolkningsbarhet, förklarbarhet av AI-algoritmer samt förbättrad transparens av AI-metoder. Det finns i dagsläget begränsad kunskap och forskning kring hur människor tänker om AI. Författarna betonar vikten av att förstå hur det mänskliga sinnet uppfattar AI för att kunna bygga upp pålitligheten och acceptansen kring systemen. I författarnas studie undersöker de människors implicita och explicita attityder till AI-system. Studiens resultat visar att majoriteten av de respondenter som var delaktiga i undersökningen som ansåg sig ha en positiv syn på AI visade sig ha en implicit negativ attityd. Detta förklarar författarna grundar sig i att användares attityder till AI påverkas av omedvetna och medfödda fördomar, vilket i sin tur kan leda till allmänna konsekvenser för samhället. Författarna beskriver att människors negativa implicita attityder till AI både kan motverka eller försena viktig utveckling av ny teknologi som kan förbättra livskvaliteten i framtiden. Genom att övervinna dessa fördomar och tankar hos människor kan förtroendet för ny teknik och AI istället öka vilket kan leda till positiva konsekvenser för samhället. Författarna förklarar att det på ett sätt kan vara farligt att minimera människors implicita negativa attityder kring AI då dessa fungerar som ett skyddande värde för människor och samhället. De bygger upp en försiktighet vilket minskar chanserna för skadligt missbruk av den nya teknologin som AI utvecklingen kan medföra. Det kommer i framtiden vara lättare och säkrare att hjälpa människor att övervinna de fördomar och den implicita negativa attityd människor har till AI på grund av att en reglering kommer framställas. Regleringen kommer ta hänsyn till både riskerna och fördelarna med AI vilket kan hjälpa människor att öka deras förtroende och acceptans för den nya teknologin (Fietta et al. 2022).

Kelley et al. (2021) undersöker människors tankar om AI inom fyra olika huvudområden spännande, användbart, oroande och futuristiskt. De respondenterna som deltog i undersökningen är utspridda bland åtta olika länder och man kan se att det både finns likheter och skillnader kring hur människor tänker om AI beroende på geografisk tillhörighet. Ser man till de fyra huvudområden författarna har djupdykt i deras undersökning kan man läsa av att 18,9 procent av respondenterna känner en positiv och spännande känsla bakom utvecklingen av AI. 12,2 procent av respondenterna känner en användbarhet i AI och tror starkt på att teknologin kommer vara hjälpsam och assistera människor i deras sätt att utföra uppgifter. När det kommer till oroligheterna kring AI upplever 22,7 procent av respondenterna negativa känslor, oro och rädsla kring utvecklingen. Till sist har vi en grupp som motsvarar 24,4 procent av respondenterna som varken har en positiv eller negativ attityd till AI men kopplar AI till ett avancerat framtidsperspektiv. Den generella slutsatsen författarna tar fram genom analys av deras resultat är att majoriteten av människorna i undersökningen tror att AI kommer ha en betydande inverkan på samhället i framtiden. Resultatet visar även att människorna i alla de åtta länderna har en större procentuell andel

som tror att AI kommer till störst del ha en positiv inverkan på samhället än en negativ inverkan (Kelley et al. 2021).

5. Resultat

Inledningsvis i detta kapitel presenteras varje respondent samt hur intervjuerna har genomförts. Därefter sammanställs resultatet och informationen från de semistrukturerade intervjuerna som har genomförts. Resultatet presenteras utifrån de respondenter som varit delaktiga i denna undersökning. Underrubriker har tagits fram och är utvecklade utifrån studiens forskningsfrågor för att säkerställa att dessa besvaras.

5.1 Respondent 1 - Olle Häggström

5.1.1 Osäkerheten kring AI-utvecklingen

Frågorna om hur AI och mer specifikt AGI kommer påverka samhället har intresserat Olle under många år. Under de senaste åren, i samband med corona-pandemin, har utvecklingen accelererat i en takt så det inte längre finns någon rimlig tvivel om att AGI-teknologin är möjlig. En teknologi som tros skulle ta minst 20 till 30 år, kanske 50 år eller till och med 100 år att utveckla kan komma att finnas mycket tidigare än vad mänskligheten tänkt sig påstår Olle. När AGI är skapat står vi inför det största paradigmskiftet någonsin i vetenskapens historia. Människan har med hjälp av sin intelligens skapat sig en unikt stark position där människan dominerar denna planet. Genom utvecklingen av AGI automatiseras människans intelligens och riskerar att hamna i ett läge där människan befinner sig på en andra plats bland varelser med högst intelligens.

Olle förklarar att det behövs en inbromsning av utveckling av AI för att utvecklarna ska kunna säkerställa modellernas säkerhet och lägga kraft på AI Alignment. Förslaget som organisationen Future of Life tar upp i sitt öppna brev om en 6 månaders paus tror Olle inte på, men det är ett steg i rätt riktning. Olle anser att det är nödvändigt att stoppa träningen av mycket stora AI-modeller utan någon fastställd tidsram, och detta stopp ska införas kraftfullt med internationella överenskommelser som stöds med militär styrka.

> *Det behövs ett stopp utav träning av riktigt stora AI modeller som inte har någon bortre tidsgräns och det här stoppet behöver implementeras med kraft med internationella överenskommelser med militär styrka bakom så att de som bryter mot dem vet vad som händer om man bryter mot dem. Därför de som bryter mot dessa överenskommelser sätter hela mänskligheten i fara* (Olle Häggström).

Olle tar upp att de värsta riskerna med AI behöver hanteras för att osäkerheten kring AI ska minskas. Risker som handlar om diskriminerande och etiska beslut samt hur AI kan användas för spridning av skadlig och missvisande information. Alla dessa risker är viktiga för osäkerheten med utvecklingen men den mest överhängande och viktigaste frågan är huruvida AI kommer förinta mänskligheten.

Vidare förklarar Olle att idag är proportionerna mellan antalet AI utvecklare och de som arbetar inom forskningsområdet AI-alignment inte balanserade. Fler forskare behövs inom området AI-alignment, som fokuserar på att superintelligenta maskiner har mål och drivkrafter som i tillräcklig mån prioriterar människors välfärd och värderingar. AI-alignment är ett svårt tekniskt problem där det är svårt att se en tydlig väg framåt. Att lyckas lösa AI-alignment är en nödvändighet för en säker utveckling anser Olle. Fortsättningsvis beskriver han hur de stora AI företagen konkurrerar om marknadsandelar och

marknadsdominans. Företagen önskar att undvika hinder i form av etiska överväganden och problematiken kring AI-alignment, och istället fortsätta utvecklingen framåt.

Olle tror att det finns en möjlighet att ett utopiskt samhälle kan designas där AI utför allt arbete åt oss. Ett samhälle där människan inte behöver fatta beslut då vi sett till att AI är utvecklat på ett vis att den alltid ska gynna människan. Samtidigt uttrycker Olle att det inte känns som ett idealiskt samhälle där mänskligt liv saknar något, en agenda. När utvecklingen av en AGI finns kommer den vara *wild* i samhället, med andra ord att den kommer kunna göra vad den vill, menar Olle. Han fortsätter med att AGI kommer att kunna ha viljan att släppa in en människa i beslutfattandet vid olika typer av scenarion. I vissa scenarion skulle människan endast bli ett problematiskt brus vid beslutfattande och vid andra beslut kanske det krävs en bättre mänsklig förståelse av de potentiella riskerna.

... det kan bli fantastiskt bra, det kan bli lösningen på alla våra problem och det kan bli slutet för mänskligheten om vi gör det på fel sätt (Olle Häggström).

5.1.2 Riskerna kring AI-utvecklingen

Den absolut största risken Olle ser med utvecklandet av AI är att människor skapar en AI som kommer att förinta hela mänskligheten. Olle anser att kraftfulla AI-system endast bör utvecklas när man kan garantera att de kommer ha en positiv inverkan på samhället och att riskerna med AI går att hantera. Han förklarar även att en annan risk med AI är att techbolag stressar med att släppa produkter innan de är säkra eftersom de strävar efter att vara först och bli marknadsdominerande. Detta förklarar Olle att vi kan leva med i nuläget men att desto kraftigare AI-systemen blir desto högre takt ökar riskerna. Om en utvecklingspaus inte genomförs och människor fortsätter utveckla kraftfulla AI-system som de inte förstår sig på tror Olle att riskerna kan bli obegränsade stora på bara några års sikt. Om allt dock genomförs på rätt sätt tror Olle att AI kan vara en räddare för framtiden.

Hur bra som helst. AI har potential att bli nyckeln som kan lösa alla de stora samhälls-, klimat- och naturresursfrågorna och skapa en fantastisk blomstrande framtid för mänskligheten. Men då måste allt bli rätt, vi kan inte rusa framåt som vi gör nu (Olle Häggström).

Olle är väldigt mån om de potentiella riskerna utveckling kan medföra och den paus Future Of Life föreslår i sitt öppna brev är långt ifrån tillräckligt för utvecklarna att säkerställa modellernas säkerhet och AI-alignment. En stor problematik är att utveckling sker på en global marknad med där regioner och länder tar fram olika krav och riktlinjer på hur AI ska utvecklas. Samtidigt som regioner och länder utvecklar rättsliga ramar som är säkra och pålitliga, är de rädda att hamna efter i AI-utvecklingen menar Olle. Fördomar som antar att exempelvis kinesiska företag inte kan ta ansvar för att undvika en potentiell undergång för mänskligheten ställer sig inte Olle bakom och antyder att det är viktigt att alla ledande AI-företag samarbetar.

Vi vänder ju oss till alla ledande AI företag. Och jag gillar inte den typen av fördomsfullhet där man tar för givet att inte kinesiska företag skulle kunna ta ett ansvar här för att undvika mänsklighetens undergång (Olle Häggström).

Olle förklarar att de stora AI företagens kapplöpning om att ta marknadsandelar och greppa en dominant position på marknaden leder till att de inte beaktar konsekvenserna ordentligt.

Den snabba utvecklingen gör att AI inte utvecklas på ett transparent och väl testat sätt vilket gör att AI löper större risk att manipulera och ge missvisande information för att uppfylla sina mål. Olle tar sedan upp risken att människor överskattar AI-systemens tillförlitlighet. Det är ett vanligt problem som alltid har funnits där han påpekar att hur även människor ger falsk information. Det är därför viktigt att applicera samma sunda förnuft och källkritik mot AI-systemen som vi har gentemot människor. Problematiken ställs på sin spets på grund av den snabba teknikutvecklingen, det är viktigt att systemen uppdateras snabbt för att uppfattas som pålitliga.

Olle tar upp riskerna med att AI förmodligen kommer förändra arbetsmarknaden rätt rejält i framtiden och att detta kommer leda till att många nuvarande mänskliga arbeten kommer att bli överflödiga. Han förklarar dock möjligheterna med att detta kommer leda till att nya jobb kommer att skapas. Den stora risken blir att säkerställa att människor som ska utföra dessa nya arbeten måste ha tillräcklig och rätt utbildning och omställningsmöjligheter för att kunna anpassa sig till de nya arbetsmarknaderna som AI-utvecklingen kan medföra.

Olle förklarar att en risk kring AI och framförallt Chat-GPT är att studenter kan använda den för att skriva inlämningar eller uppsatser. Risken som definieras som fusk var framförallt aktuell när Chat-GPT lanserades av olika professorer på olika forskningsinstitut. Olle menar på att diskussionen har utvecklats eftersom de flesta kommit till insyn att ifall vi inte anpassar oss till utvecklingen erbjuds en föråldrad utbildning. Det är däremot viktigt att anpassa utbildning på högskola och universitet så att dessa typer av AI-bottar framförallt bidrar till lärande istället för fusk. Olle förklarar vidare att Chalmers, där han arbetar, har framfört riktlinjer kring hur dessa typer av händelser ska hanteras. Tyngden av att dessa typer av riktlinjer är levande och är av högsta vikt eftersom riktlinjerna troligen kommer behöva uppdateras inom 6 till 12 månader. Olle menar på att riktlinjerna kommer behöva uppdateras eftersom utvecklingen går i en hastig takt och att situationen kan vara en helt annan inom en relativt snar framtid.

I intervjun förklarar Olle att en risk med Chat-GPT är dess förmåga att skapa falska referenser gentemot användaren. Det är inte helt säkert varför chatbotten inte kan leverera valida referenser men Olle för en diskussion där han tror att utvecklarna har velat undvika en typ av klipp och klistra variant av referenser. Tanken kring diskussion är att utvecklarna önskar att botten själv med dess intelligens ska kunna leverera korrekta referenser utan att behöva klippa och klistra.

Risken kring att ett flertal olika aktörer utvecklar egna AI-chatbotar och försöker ta marknadsandelar menar Olle är ett problem. Han berättar om botten som Microsoft lanserade där botten lanserades och var väldigt diskriminerande gentemot dess användare. Problematiken kring den hastiga utvecklingen är alltså just att viss osäkerhet inte är diskuterad utan att det prioriterade målet är att lansera dessa typer av tjänster så snabbt som möjligt. Olle förklarar vidare att just denna typen av AI inte är mer farlig än andra typer av sökningar på internet. Den större risken är att detta tankesätt och konkurrens mentaliteten fortsätter när mycket mer kraftfulla typer av AI lanseras. Olle menar att om inte ett flertal säkerhetspunkter har följts på ett tillfredsställande sätt kommer detta innebära en anmärkningsvärd problematik för framtidens samhälle. Den främsta oron Olle beskriver är att en tillräcklig osäker och superintelligent AI har möjligheten att ta över makten över världen utan att någonsin behöva lämna sitt kontor.

5.2 Respondent 2 - Josefin Rosén

5.2.1 Osäkerheten kring AI-utvecklingen

Josefin inleder i sin intervju med att förklara att människor varken borde eller kan stoppa den stundande AI-utvecklingen i dagens samhälle. Hon anser att människor inte behöver vara oroliga utan att man istället ska vara glad över utvecklingen då den öppnar upp för ofantliga möjligheter och fördelar. Josefin förklarar att det öppna brevet som publicerades i mars 2023 har bidragit till en onödig rädsla och skrämsel bland människor. Det är svårt att veta vad de riktiga intentionerna med brevet var menar Josefin men att det medfört en stor rädsla kring hur systemen fungerar och vad de kommer kunna utföra i framtiden. Det öppna brevet har även medfört att det blivit ett väldigt stort fokus på att AI är farligt och att AI behöver stoppas innan det är försent vilket Josefin tycker både är onödigt och beklagligt. Modellerna tränas inte på det sättet.

Det finns risker med AI såklart och dem har ju funnits hela tiden och i många år och det blir på en annan skala men då är det inte framförallt det att det ska komma en terminator och liksom ta över världen, där är vi inte idag (Josefin Rosén).

Josefin fortsätter med att beskriva att osäkerheten kring AI grundar sig i den stora okunnigheten kring ämnet i samhället. Detta är ingenting som hon anser är konstigt utan menar att allting som människor inte riktigt förstår sig på blir läskigt och då bildas en osäkerhet. Josefin bekräftar att osäkerheten kring AI är befogad men att den inte nödvändigtvis behöver vara så dramatisk som den i dagens samhälle har blivit.

Josefin tycker att det borde ligga ett större fokus på hur man ska hantera riskerna med AI och behovet av att hantera dessa risker istället för att fokusera på att AI är farligt. Josefin betonar vikten i att det i framtiden kommer finnas en reglering av AI som kommer hjälpa till att öka förståelsen och minska osäkerheten. Trots att en reglering håller på att tas fram och förhoppningsvis kommer vara färdigställd under nästa år är det fortfarande viktigt att öka förståelsen kring AI i samhället. Josefin förklarar att en av de viktigaste faktorerna är att man sprider kunskapen om hur verktygen används på rätt sätt så att människor blir medvetna om de risker som finns, vad man kan lita på och vad man inte kan lita på när det kommer till AI.

Vidare förklarar Josefin att utveckling av transparens och förklarbarhet kring hur AI-systemen fungerar är viktiga faktorer för att öka säkerheten och förtroendet för systemen. Detta är något som Josefin förklarar var dåligt utvecklat för ett antal år sedan men något som forskare arbetat väldigt mycket med under de senaste åren. Detta har resulterat i att man kunnat ta fram metoder som arbetar med att göra AI-modellerna förklarbara och detta är något som forskare arbetar med att ständigt utveckla. Josefin tar upp OpenAI som exempel och förklarar att hon tycker det är konstigt att de inte har gått ut med hur allt bakom det fungerar med tanke på deras syfte från början. Därför är det också svårt att säga hur transparent det egentligen är. Josefin beskriver att det är viktigt att det finns transparens i hela kedjan. Det finns metoder för att förklara och förstå algoritmerna och modellerna men det är lika viktigt att förstå vart datan kommer ifrån, vad man gjort med datan som man tränat på, hur man manipulerat eller kompletterat datan, hur besluten tas baserat på resultatet som kommer från modellerna och hur man väljer att hantera det utfall som modellen genererar. Detta förklarar Josefin resulterar i att man får transparens i hela kedjan och när man vet hur något fungerar blir det även lättare att lita på. Detta tillsammans med regleringen anser Josefin kommer öka säkerheten och förtroendet för AI-användning i samhället.

Vet man att det ställs krav på dokumentation, rapportering och transparens osv på dem kraven som finns så vet man också att det är tryggt att använda det som har hög risk i samhället och det är ju hela förhoppningen med AI, att man ska kunna känna sig trygga med att använda den AI som finns i samhället, det är min förhoppning, att det faktiskt ska bidra till det (Josefin Rosén).

Det pratas mycket om förväntningarna och möjligheterna att AI i framtiden kommer kunna fungera helt autonomt men detta tror inte Josefin är fallet utan hon tror att det alltid kommer behövas en human in the loop. Även fast hon tror att det kommer gå att helautomatisera och låta AI ta hand om mycket i framtiden anser hon att det alltid kommer krävas en övervakning av systemen. Övervakningen handlar inte om att bevaka AI så den inte springer iväg och utför kritiska saker för det är inte fallet. Utan att det krävs övervakning på grund av att man tränar datan på den verklighet som vi är i exakt nu för att det ska vara så relevant som möjligt, men att verkligheten kan ändras drastiskt och detta är inte säkert att AI kan ta hänsyn till. Josefin tar upp pandemin och Ukraina kriget som exempel på när mycket ändrades över bara en natt och att detta medförde att många modeller och data blev inaktuella på väldigt kort tid. Därför anser Josefin att det alltid kommer behövas en human in the loop, för att ha möjligheten att gå in och se att allting verkligen blir rätt. Det handlar framförallt om att kontrollera att modellen hänger med och ifall den glider från korrekt linje måste den kunna tränas om, så modellen verkligen är anpassad efter den verklighet vi har just nu. Detta är något som Josefin beskriver som väldigt viktigt när det kommer till AI.

5.2.2 Riskerna kring AI-utvecklingen

I intervjun med Josefin förklarar hon att de risker som finns inom AI gäller samtliga system. Det handlar om misinformation, bias eller diskriminering gentemot individer inom samhället. Josefin förklarar även att risken är att dessa kraftfulla verktyg hamnar i fel händer. Något som har ett tänkt bra ändamål kan missbrukas, detta förklaras inte unikt för AI utan gäller vilken typ av IT-lösning eller verktyg som helst.

Dem riskerna som jag ser är de risker, det finns risker med alla AI-system, att de hamnar i fel händer och att det blir någon acceleration av misinformation eller acceleration utav bias eller diskriminering osv (Josefin Rosén).

En lösning för möjliga risker förklaras med hög transparens och ett stabilt government system för att uppnå kontroll över utvecklingen. API:er som är tillgängliga eller åtminstone är åtkomstbara ses som en stor risk. Problematiken ligger då i att det är möjligt att komma åt och bygga egna system. De egenkomponerade systemen är då en risk eftersom de kan användas till ändamål de inte var tänkta att användas för. Josefin förklarar även att genom användning av API:er är det troligt att individer inte kommer att insiktsfullt förstå vad som sker i själva grundmodellen, vilket leder till en låg förståelse kring hur modellerna egentligen fungerar.

Vidare i intervjun förklarar Josefin att användare eller utvecklare inte naivt får acceptera samtliga svar eller resultat som AI genererar. Hon förklarar att det krävs en viss kritisk syn för framför allt utvecklare. Effekten av den kritiska synen ser till att parametrarna ständigt utvecklas för att resultatet ska uppnå ett så optimalt svar eller resultat som möjligt. Josefin påstår även att AI endast är ett verktyg eller komplement, det är alltså nödvändigt med mänsklig interaktion. Josefin fortsätter i intervjun att beskriva att trots att AI är ett väldigt

kraftfullt verktyg krävs det att människor fortsätter tänka för sig självt. Hon drar en parallell gentemot att skriva uppsats och förklarar att det går att få väldigt mycket bra information från ett verktyg som Chat-GPT men att som forskare krävs det även vid seminarier eller liknande att kunna motivera och försvara sin undersökning. Att förklara varför en viss metod har valts och argumentera kring det kan inte en AI göra på ett liknande sätt som en människa. Risken finns alltså att bli bekväm när man använder AI som Chat-GPT eller liknande och glömmer att tänka och resonera kring ämnet själv.

En diskussion förs även kring Chat-GPT där Josefin menar på att det är ett utomordentligt verktyg för nya tankar, idéer eller infallsvinklar. Josefin förklarar att hon använder chatbotten i sitt arbete men att risken ligger i vart botten hämtar informationen från. Chat-GPT upplyser inte användaren om referenser utan ger endast ett svar, därför påpekar Josefin att det bör ligga i ens intresse att kontrollera om informationen är korrekt.

Jag använder det verkligen, i jobbet, det gör jag. Men man måste kolla referenserna, man måste kolla upp så att det stämmer (Josefin Rosén).

Gällande risken kring arbetslöshet eller att AI kommer ersätta jobb i framtiden tror Josefin att det är troligt, eller åtminstone att ett antal typer av arbetsuppgifter kommer att effektiviseras. Hon förklarar däremot att historiskt sett när en revolutionär utveckling skett har samhället utvecklats och nya jobb och arbetsuppgifter skapats. Josefin ser utvecklingen som något positivt och anser att möjligheten att utföra saker snabbare med bättre resultat inte kommer öka arbetslösheten.

Jag tror att det kommer säkert ersätta en del arbeten, i alla fall hjälpa vissa arbeten att bli mycket mer effektiva. Jag tror inte man behöver vara oroliga över det för historiskt sett har man sett när, när det kommit nya möjligheter då har också arbetsuppgifterna utvecklats så det brukar ju snarare vara en revolutionär fördel för att samhället ska gå framåt. Det är många som säger att man ska hitta meningsfulla mänskliga uppgifter och så kan det ju också va men jag tycker vi ska se en utveckling att vi avancerar och får liksom en helt annan output att göra saker snabbare och jag tror att det bidrar till utvecklingen jag tror inte det kommer bidra till att folk blir arbetslösa (Josefin Rosén).

När Josefin fick svara på frågan *Vad är den värsta tänkbara konsekvensen av dagens utveckling?* förklarar hon att det snarare handlar om att personlig data eller privacy breaches exponeras. Utöver att personlig data exponeras förklarar Josefin återigen att en av de största riskerna är att det hamnar i fel händer. En risk som även tas upp förklarar Josefin inte troligen är en risk i Sverige utan i länder där omvärldsbevakning är begränsad. Hon påstår då att i länder som inte har den öppenhet Sverige har finns en större risk för spridning av falsk information. Josefin nämner för att inte detta ska drabba individen krävs en öppenhet, en transparens och ett typ av etiskt filter för att säkerställa outputten. Josefin förklarar även i intervjun att hon inte tror AI kommer bli farligt för mänskligheten.

Slutligen i intervjun förklarar Josefin att även om utvecklingen går väldigt snabbt framåt nu och att det finns en otrolig uppsida kring AI-utvecklingen är det lätt att för liten vikt läggs kring de möjliga risker som kan medföras. Balansen anser Josefin är viktig och nödvändig för att tillförlitligheten kring AI ska vara hög. AI med hög tillförlitlighet kommer sedan att ge mänskligheten en stor fördel och då förbättra lösningar inom områden som sjukvård och klimat. Avslutningsvis förklarar Josefin att en reglering krävs för att individen ska känna sig säker med användningen av AI.

... och att man jobbar med att det faktiskt finns. Kommer en reglering så att folk kan känna sig trygga, då kan man också komma framåt för att finns det "Trust" finns det också möjlighet till innovation (Josefin Rosén).

5.3 Respondent 3 - Mathias Lanner

5.3.1 Osäkerheten kring AI-utvecklingen

Mathias inleder i sin intervju med att berätta och förklara hur han använder och tillämpar AI i sitt dagliga arbete. Han berättar att det som har utvecklats inom AI från och med november 2022 fram till nu är en utveckling som inte har hänt under hans 25 år i arbetslivet.

Det är en teknikutveckling som är helt sjuk just inom det här området som alla på något sätt kan ta till sig för vem som helst kan skriva en fråga och få fram något som är skitbra, detta är ren magi (Mathias Lanner).

Mathias fortsätter med att förklara att vi inte behöver vara oroliga över den snabba utveckling som AI idag har i samhället men att det är viktigt att vi är försiktiga. Den enda anledningen till att vara oroliga för utvecklingen är om forskarna inte lyckas utveckla någon kontrollfunktion för systemen. Han betonar vikten med att kloka regelverk måste utvecklas och tas i bruk och inte bara från EU utan dessa regelverk måste vara världsomspännande. Det är viktigt att man genom klokskap hittar någon form av kontroll över systemen som inte är innovationsdödande för att detta ska bli så bra som möjligt och för att minska osäkerheten menar Mathias. Han tycker heller inte att den stundande utvecklingen av AI är utom kontroll för tillfället. Det har kommit nya versioner den senaste tiden som talar om vilken data och information som modellerna tränas på vilket stärker tillförlitligheten för systemen. Han förklarar hur man kan minska osäkerheten ytterligare och stärka förtroendet genom att utveckla en funktion som verifierar utdatan som AI genererar. Det handlar om att kunna göra en validering av det svar som AI tar fram för att kunna kontrollera att utdatan är korrekt samt följer etiska normer och riktningar menar Mathias.

Mathias tror inte att det kommer gynna samhället om utvecklingen stoppas eller pausas då detta kan hindra utveckling av teknologi som kan skapa stora möjligheter och positiva förändringar för både människor och samhället i framtiden. Då Mathias anser att dagens utveckling är under kontroll beskriver han att det bara är farligt att införa ett förbud. Han tror inte att detta är något som bara kan förbjudas och sen kommer detta att följas. Han har svårt att tro att företag och forskare som idag arbetar med att utveckla teknologin kommer sluta med detta oavsett om ett förbud införs. Det kommer även vara svårt eller helt omöjligt att kontrollera att detta följs och därför menar Mathias att ett förbud inte kommer göra något bra. Det kommer bara leda till ytterligare konsekvenser, poängterar han.

Förbjuda kommer inte att göra något bra, för det kommer inte att kunna gå att förbjuda nåt sånt här (Mathias Lanner).

För att öka säkerheten och förtroendet för AI-systemen förklarar Mathias att det är viktigt att öka kunskapen om teknologin i samhället. Det handlar inte bara om att öka kunskapen om det som är positivt med systemen utan även det som är negativt och framförallt kunskapen om att systemen kan göra fel, poängterar Mathias. Han anser att de människor som använder systemen bör utbildas och att detta borde göras i ett tidigt stadium, exempelvis redan i skolan. Detta för att öka kunskapsnivån kring teknologin, hur systemen ska användas och vilka risker som finns med systemen. Det resultat som genereras är inte alltid rätt, det behöver finnas en förklaring till hur detta har framställts och att detta inte alltid kommer vara sanningen utan att det kommer från en beräkning av den text som man skrivit in. Mathias förklarar att det går att öka säkerheten med AI ytterligare om det exempelvis hade utvecklats en relevansfunktion för den utdata AI genererar. Han föreslår en framtida funktion som mäter relevansen i procent av utdatan för att människor själva sen ska kunna avgöra hur mycket det går att lita på resultatet, detta tror Mathias hade ökat tillförlitligheten med systemen. Det sista Mathias tar upp med hur man kan öka säkerheten med AI-systemen är att skapa någon övergripande tillsynsmyndighet som övervakar utvecklingen. Han anser att det är en viktig faktor att det skapas en kontroll för att övervaka att människor inte utvecklar något som är farligt med hjälp av teknologin. Mathias föreslår att en lösning hade kunnat vara att införa en regulatorisk process för att få använda den teknologi man utvecklar, liknande den som redan finns i läkemedelsindustrin. Då hade det kunnat fungera så att den teknologi som utvecklas måste gå igenom en kontrollfunktion som ett antal steg, prövningar och sen godkännas för att få användas. Detta tror Mathias hade haft en betydelsefull mening för att öka säkerheten och förtroendet för systemen.

När det kommer till chansen att AI ska kunna fungera helt autonomt i samhället beskriver Mathias att utvecklingen inte är där idag men att forskare säkert kommer lyckas utveckla detta i framtiden. Trots att det finns en möjlighet till att helautomatisera med hjälp av AI anser Mathias att AI inte bör bli helt autonomt utan att det alltid ska finnas en interaktion med människan. Han förklarar att människor kan tänka och bete sig på ett sätt som systemen aldrig kommer kunna göra och därför anser han att det är lika viktigt att dra nytta av båda aspekterna.

Jag tror på människorna, att det alltid måste finnas en symbios tror jag, absolut (Mathias Lanner).

5.3.2 Riskerna kring AI-utvecklingen

I intervjun med Mathias förklarar han risken att undervisningen inte hänger med i utvecklingen och berättar att ett flertal lärosäten försökt förbjuda AI eller Chat-GPT. Mathias menar att olika typer av språkmodeller behöver anpassas till undervisningen istället för att det ska förbjudas. Han förklarar även att även om mycket information är relevant finns möjligheten att det är felaktigt. Han ser även den nya tekniken som ett potentiellt innovationsdödande verktyg. Mathias menar att det är enkelt för studenter eller barn att använda verktyget utan att själva tänka kring ämnet. Han berättar i intervjun att dessa typer av verktyg inte främjar kreativitet och tankekraft utan försämrar den. Mathias är däremot delad kring användning av språkmodeller eftersom han anser att det är ett verktyg som på ett

smidigt sätt ger resultat och assisterar individen. Han anser även att ett AI-förbud också skulle vara innovationsdödande.

Det är bara att formulera en fråga. Sen får man fram ett svar som man på nåt sätt tycker, det främjar inte egen tankekraft. Det kan vara innovationsdödande just för de yngre människorna (Mathias Lanner).

Mathias förklarar att han tror att ett antal företag och yrken kommer att försvinna med den nya tekniken. Han förklarar att det är naturligt att med tiden kommer ett flertal jobb försvinna medans ett par nya kommer att tillkomma. Utvecklingen kommer även kräva att kunskapsnivån generellt behöver öka bland människor, beskriver Mathias. Människor kommer alltid vara sysselsatta och de som blir ersatta med AI kommer hitta något annat. Han är generellt sätt inte orolig över arbetsmarknaden eftersom innovation skapar nya typer av tjänster och arbetsuppgifter. Mathias förklarar även en risk kring att företag för in kod till AI utan att förstå konsekvenserna eftersom det är smidigt och lättillgängligt. Den datan som matas in kan innehålla konfidentiell information som exempelvis Chat-GPT eller liknande verktyg får tillgång till. Mathias menar att om privat data matas in äger AI-verktygen den informationen och kan göra vad de vill med den, det förklarar Mathias som en risk.

Så har det på nåt sätt varit att om vi ersätter jobb med saker så kommer människorna göra andra saker. Jag är inte så orolig (Mathias Lanner).

Möjligheten kring att det kan bli felaktiga resultat genom AI-system ser däremot inte Mathias som en stor risk. Han förklarar att det är viktigt att poängtera för användaren att svaren AI inte alltid stämmer utan att det är en prediktion AI gör. Han diskuterar att en “relevansscore” kan vara en hjälpsam funktion för utvecklingen där användare får betygsätta outputten. Det tror han även kommer hjälpa på så sätt att utvecklingen inte skenar iväg.

Risken kring att AI-utvecklingen i framtiden kan hota mänskligheten tror Mathias inte på. Däremot ser han riskerna kring om dessa kraftfulla verktyg hamnar i fel händer kan olyckliga situationer uppstå. Ett exempel som gavs under intervjun var kapning av kärnvapenterminaler, han förklarar däremot att vi inte är där nu inom utvecklingen.

5.4 Respondent 4 - Mattias Ohlsson

5.4.1 Osäkerheten kring AI-utvecklingen

Mattias berättar i sin intervju att han redan från början av AI-utvecklingen har varit insatt i området och tycker att det är ett väldigt intressant ämne att forska kring. Han beskriver hur AI kan användas för att göra fantastiska saker, framförallt de stora språkmodellerna och hur dessa kan hjälpa oss i samhället. Han bekräftar att det finns en osäkerhet kring den stundande utvecklingen av AI men anser att forskarna har utvecklingen under kontroll och är därför inget människor behöver vara oroliga över i dagsläget. Själva utvecklingen ser Mattias alltså inga risker med utan beskriver istället sin höga förväntan av att nå full artificiell intelligens i samhället. Mattias förklarar dock att det finns en osäkerhet som människor kan behöva vara oroliga för men det handlar inte om utvecklingen utan mer om riskerna att exempelvis teknologin används på fel sätt.

… jag tycker att det är en spännande utveckling och jag ser inga farhågor i utvecklingen vad gäller forskning kring det här (Mattias Ohlsson).

När det kommer till den 6 månaders utvecklingspaus som föreslagits tror Mattias att detta inte kommer leda till något bra utan bara hämma utvecklingen. Han tror inte att AI kommer ta över mänskligheten eller att det kommer bli något slags terminator scenario som det idag finns oro för i samhället. Mattias poängterar att det finns vissa delar av teknologin som kommer att komma efter den snabba utvecklingen och inte hänga med. Detta är något som ofta förekommer med ny teknik men kommer att balanseras upp desto längre tid man forskar kring ämnet, förklarar Mattias. Han beskriver att en utvecklingspaus hade kunnat vara effektiv för att sätta sig ner, diskutera och öka förståelsen kring utvecklingen av teknologin men då hade det krävts att detta införs globalt. Detta är anledningen till att Mattias anser att en utvecklingspaus inte hade varit effektiv då han förklarar att det hade varit helt omöjligt att få alla att pausa utvecklingen framför allt de med onda avsikter. Han förklarar även att det hade varit helt omöjligt att kontrollera att alla tar sitt ansvar och följer en potentiell utvecklingspaus och ser därför att det hade medfört mer nackdelar än fördelar.

För att det ska vara effektivt behöver alla göra det. Jag tror det är helt omöjligt att få alla att göra det. Framförallt inte dem som har onda avsikter, kommer naturligtvis inte halta en sån utveckling (Mattias Ohlsson).

Mattias är skeptisk till om AI någon gång kommer kunna fungera helt autonomt i samhället och ser inte riktigt hur det skulle kunna fungera. Han beskriver att det framförallt är människor i samhället som driver denna utveckling framåt på det sättet att det efterfrågas något som kan byta ut det mänskliga arbetet och därför utvecklas saker på ett visst sätt. Mattias tar upp självkörande bilar som ett exempel, där människor efterfrågar någonting som till grund har en tanke att fungera helt autonomt. Mattias beskriver att man redan idag kan hitta och se autonoma funktioner i små subsystem men att skala upp detta till större saker och ta bort den mänskliga interaktionen både kan leda till risker och osäkerhet.

5.4.2 Riskerna kring AI-utvecklingen

Mattias förklarar i intervjun att utvecklingen är något positivt men att ny teknik kan användas på fel och direkt skadligt sätt gentemot samhället och individen. AI kan användas för att exempelvis syntetisera tal vilket får folk att tro att de pratar med någon de kan lita på, vilket inte är fallet. Mattias menar att det finns aktörer som använder eller kommer att använda teknologin med fel intentioner. Risken kring att AI kommer att agera på egen hand och ses som ett hot mot mänskligheten tror Mattias väldigt lite på. Han ser en större risk att AI kan hamna i fel händer och att teknologin används eller utvecklas på ett skadligt sätt.

Jag är inte orolig för domedags tänket men jag är orolig kring att man utnyttjar tekniken i dåligt syfte (Mattias Ohlsson).

Vidare i intervjun förklarar Mattias att den legala sektorn kommer att möta svårigheter kring den nya teknologin eftersom de behöver inse hur de ska agera inom en AI-värld. Han förklarar att de behöver förstå hur de ska hantera legala ärenden kopplat till AI samtidigt som att ny innovation alltid har påverkat den legala sektorn.

Mattias förklarar risken att AI kommer ersätta jobb men ser det som naturligt. Han förklarar att innovation historiskt ersatt arbeten men att nya arbeten har skapats. Mattias anser att balansen mellan nya och gamla jobb kommer att förbli stabil. Ett exempel Mattias tar upp är när datorerna kom, då var många människor oroliga att deras arbete skulle ersättas av datorerna men det blev istället ett effektivt verktyg i deras arbete.

Ytterligare en risk som Mattias tar upp är risken kring bias. Han förklarar att denna risk existerar eftersom modellerna matas med enorma mängder text. Den data eller text modellerna tar till sig kan vara vinklad vilket medför att modellerna i ett senare stadie kan ge en bias output. Mattias menar att det inte finns någon mystik kring hur AI fungerar utan att det är enkelt att förstå modellerna och varför de genererar ett visst svar. Bias är helt enkelt en risk av att AI hanterar enorma mängder text.

Det är alltså en konsekvens av all den textmassan modellen har sett att den då kan bli bias.
(Mattias Ohlsson).

6. Analys och diskussion

I detta kapitel presenteras en analys av resultatet från intervjuerna med AI-experterna, i samband med det teoretiska ramverket som bygger på dokumentation och tidigare litteratur. Genom att identifiera överlappningar eller skillnader mellan den insamlade datan och litteraturen möjliggörs diskussion att utmana och utvidga den teorin som bygger på tidigare forskning och dokumentation. Målet är att stärka och fördjupa studiens resultat.

Kapitlets avsikt är att analysera och diskutera befogenheten av osäkerheten kring utvecklingen av AI, baserat på experters synpunkter och befintlig forskning. Vidare kommer kapitlet att presentera de risker som identifierats av AI-experter. Genom att integrera experters synpunkter, befintlig dokumentation och forskning, kommer rapporten bidra till en mer omfattande förståelse kring osäkerheten och riskerna med utvecklingen av AI.

Analysen delas upp utifrån undersökningens två forskningsfrågor för att på bästa sätt säkerställa att dessa frågor besvaras.

- *Är osäkerheten med AI-utvecklingen befogad?*
- *Vilka risker kan uppkomma med AI:s fortsatta okontrollerade utveckling utan mänsklig förståelse?*

Är osäkerheten med AI-utvecklingen befogad?

Bedömningen av osäkerheten kring AI-utvecklingen är en komplex fråga som kan vara svår att genomföra på ett konkret sätt. Wu och Shang (2020) beskriver att bristande information, otillräcklig förståelse och svårighet att differentiera mellan alternativ är tre faktorer som orsakar osäkerhet. Den vanligaste orsaken är enligt författarna ofullständig information som handlar om hur AI samlar in information. Tillräcklig och adekvat information är ett krav för att resultatet AI konstruerar och genererar ska vara tillförlitligt och inte skapa osäkerhet, förklarar författarna. För att AI ska fungera på ett tillfredsställande sätt måste den insamlade datan vara av hög kvalitet och korrekt karakteriserad. AI har svårt att korrekt tolka och hantera data beroende på dess karaktär. Bristande förmåga att karakterisera om datan är objektiv eller subjektiv samt primär eller sekundär, leder till missvisande eller felaktiga resultat, förklarar författarna (Wu & Shang, 2020). Tamboli (2019) lyfter också upp denna faktor som en osäkerhet och menar att tillförlitligheten hos AI-system beror på den träningsdata den tillhandahåller. Innehåller träningsdatan ofullständig information finns det risk att AI genererar felaktig eller skadlig information (Tamboli, 2019). Olle menar att det är nödvändigt att öka forskningen inom AI-alignment för att minska osäkerheten. AI-systemen måste förstå vilka mål som är önskvärda vilket kräver att informationen är precis och fullständig. För att stärka förtroendet och minska osäkerheten anser Mathias att det bör utvecklas en funktion som kan verifiera det resultat AI genererar. Mathias menar att det är en viktig del att öka kunskapen om både de positiva och negativa aspekterna av AI-system i samhället för att minska osäkerheten och öka förtroendet. Josefin resonerar på liknande sätt och påpekar att det är viktigt med transparens och det krävs en kritisk syn på de svar eller resultat AI genererar. AI ska användas som ett verktyg eller komplement som ska ge användaren nya idéer, tankar och infallsvinklar menar Josefin. Josefin påpekar att AI-system tränas på data som speglar den aktuella verkligheten vi lever i för att vara relevant, men att verkligheten kan förändras snabbt och drastiskt. Detta innebär att AI-system riskerar att bli inaktuella väldigt snabbt när det sker snabba förändringar i verkligheten. Olle uttrycker också

oro för detta och menar att systemens pålitlighet påverkas av hur snabbt systemen uppdateras i förhållande till aktuella händelser.

Wu och Shang (2020) förklarar hur osäkerhet skapas om människor har en otillräcklig förståelse av hur beslutsproblem fungerar och vilket sammanhang de existerar i, det kan vara riskfyllt att överlåta otydliga eller skadliga processer och strukturer till AI-system. Författarna menar att detta kan leda till att AI-system räknar ut optimala lösningar som inte löser rätt problem eller till och med orsakar skada (Wu & Shang, 2020). Tomsett et al (2020) förklarar att det är framför allt när AI genererar autonoma lösningar som den mänskliga förståelsen minskar för systemen (Tomsett et al. 2020). Regler och lösningar som AI skapar autonomt kan vara omöjliga för en människa att förstå och därför skapas en osäkerhet (Bubeck et al. 2023). En tänkt lösning på detta problem förklarar Tomsett et al (2020) är att inte tillåta AI-modeller att generera autonoma lösningar utan istället uppnå mänsklig förståelse genom att inkludera en human in the loop under hela vägen fram till att resultatet är säkerställt (Tomsett et al. 2020). Olle anser inte att detta är nödvändigt och tror istället att det finns en möjlighet att bygga ett samhälle där AI kan ersätta mycket av dagens arbete till autonoma lösningar. Han tror även att det finns en möjlighet i framtiden att utveckla en AI som alltid har till grund att gynna människan. Trots detta förklarar Olle att han tror att när utvecklingen av AGI finns i samhället kommer denna teknologi vara wild och kunna göra vad den vill. AGI kommer ha viljan att kunna släppa in människan i olika scenarion men kommer i många fall endast resultera i ett problematiskt brus för teknologin menar Olle. Josefin anser att man med hjälp av AI kommer kunna helautomatisera samhället och låta AI ta hand om mycket i framtiden. Hon förklarar dock att lösningen inte är att AI får fungera helt autonomt utan att det alltid kommer krävas en human in the loop för att ständigt övervaka systemen. Det går inte att lita fullt ut på teknologin utan det kan bli fel menar Josefin, därför kommer det alltid krävas en human in the loop för att kunna kontrollera att modellerna hänger med och se till att de alltid är anpassade till den verklighet vi har just nu. Mathias har liknande tankar och inställning som Josefin när det kommer till detta område. Han förklarar att forskare inte kommer att ha några problem med att utveckla autonoma AI-lösningar i framtiden men tycker att det alltid bör finnas en interaktion med människor. Detta eftersom att han anser att oavsett hur långt utvecklingen går kommer systemen alltid kunna göra fel. Systemen kommer aldrig heller kunna tänka och bete sig exakt som en människa och därför menar Mathias att det alltid kommer krävas en interaktion då det är lika viktigt att dra nytta av båda aspekterna. Även Mattias har en skeptisk syn på om AI någon gång kommer kunna fungera helt autonomt i ett större sammanhang. Han förklarar att man idag kan se delvis autonoma funktioner i små subsystem men ser inte hur det skulle kunna fungera i en större kontext. Han beskriver att den mänskliga interaktionen är viktig och att osäkerheten och riskerna kan öka ifall den mänskliga interaktionen utesluts. Detta är även en av de faktorerna Wu och Shang (2020) tar upp som kan leda till osäkerhet. Författarna menar att de beslut AI ställs inför kan sakna en tydlig preferens eller skillnad mellan de tillgängliga alternativen vilket kan orsaka osäkerhet eftersom det kan vara svårt att förutspå hur besluten kan komma att påverka framtida händelser (Wu & Shang, 2020).

Future of life (2023) har kommit med ett förslag om en 6 månaders utvecklingspaus som ett stort antal AI experter har skrivit under. Syftet med denna utvecklingspaus är att pausa utvecklingen av AI för att se till att osäkerheten och riskerna med AI inte ökar (Future of life, 2023). Olle är en av de personer som skrivit under detta brev och ser utvecklingspausen som något absolut nödvändigt då han anser att utvecklingen är utom kontroll. Olle anser att en 6 månaders paus är ett steg i rätt riktning för att utvecklarna av AI ska kunna säkerställa och öka modellernas säkerhet. Han tror dock att en 6 månaders paus inte kommer räcka utan att

det kommer krävas ett längre stopp utan någon fastställd tidsram då utvecklarna behöver tid för att utveckla AI alignment. Vilket Olle förklarar är en av de viktigaste och mest avgörande faktorerna för att kunna fortsätta utvecklingen av AI med säkerhet. Mathias tankar kring utvecklingspausen skiljer sig mycket från Olles. Mathias anser att utvecklingen är under kontroll och att det därför är både onödigt och farligt att införa ett förbud. Han ser även det som något väldigt negativt att hindra eller skjuta upp den framtida utvecklingen av teknologin som kan skapa stora möjligheter och positiva förändringar för människor och samhället. Mathias tror även att om ett förbud införs hade detta inte följts och det hade även varit svårt för samhället att kontrollera, vilket därför endast hade medfört konsekvenser. Mattias delar liknande tankar om att en eventuell utvecklingspaus hade medfört mer nackdelar än fördelar. Mattias beskriver att han tror att en utvecklingspaus endast hade kunnat vara effektiv om detta införs globalt och om alla följer detta, något som han dock ser som helt osannolikt. Josefin har valt att inte skriva under brevet då hon varken tycker att människor borde eller kan stoppa den stundande AI-utvecklingen precis som Mathias. Hon anser att brevet endast har bidragit till en onödig rädsla och skrämsel bland människor när utvecklingen egentligen borde bidra till positivitet då den öppnar upp för ofantliga möjligheter och fördelar som man för några år sedan aldrig trodde var möjliga. Josefin tycker att det både är onödigt och beklagligt att forskare och experter bygger upp en farlig bild av AI. Josefin bekräftar att osäkerheten kring AI är befogad men att den inte behöver vara så dramatisk som den i dagens samhälle har blivit. Lägger man istället det största fokuset på behovet att hantera riskerna med AI, hur man ska hantera dessa risker och ökar förståelsen för hur AI fungerar i samhället, förklarar Josefin att osäkerheten inte behöver finnas. Hon poängterar att den framtida regleringen av AI kommer hjälpa till att sprida kunskapen om hur verktygen ska användas på rätt sätt, vilka risker som finns, vad man kan lita på och inte, vilket kommer öka förståelsen ytterligare och minska osäkerheten kring AI i samhället. Fietta et al (2022) betonar också vikten av regleringens sätt att kunna stödja utvecklingen framåt. De kommer fram till i sin undersökning att användares attityder till AI och ny teknik påverkas av omedvetna och medfödda fördomar vilket ofta är förknippat med negativa attityder. Författarna beskriver att regleringen kommer vara ett stöd för att få människor att kunna övervinna de fördomar och negativa attityder som de idag har för teknologin vilket kommer öka förtroendet och acceptansen för teknologin (Fietta et al. 2022).

Vilka risker kan uppkomma med AI:s fortsatta okontrollerade utveckling utan mänsklig förståelse?

En risk med den hastiga utvecklingen är att missvisande, felaktiga eller kränkande resultat kan genereras. Det undermåliga resultatet kan bero på att datamaterialet, insamlingen av data eller hur utvecklingen av systemet har genomförts (Morales-Forero et al. 2023). Risken för resultat som bidrar till en olycklig situation är hög då ett flertal steg kräver fakta och en värderingsmässig neutralitet (Ryan, 2020). I OpenAI rapporten (2023) påvisades ett flertal framgångar men även begränsningar. Det visade sig att outputten i vissa fall var fördomsfull eftersom modellen inte kunde skilja på säker respektive osäker data (OpenAI, 2023). Därför krävs nya nivåer och incitament av konfidentialitet, tillsammans med försäkringar om integritet för att behålla kontrollen (Bubeck et al. 2023). Olle förklarar att den snabba utvecklingen leder till att transparensen och testningen inte är tillräcklig, vilket leder till att AI löper större risk att manipulera och generera missvisande information. Det är därför viktigt enligt Olle att utveckla AI med försiktighet och förståelse kring vad som är missvisande eller inte. AI-system är ännu inte tillräckligt källkritiska vilket är en stor risk då systemen kan generera resultat på ett självsäkert sätt som i själva verket är felaktigt och möjligen även kränkande. Olle förklarar även att Chat-GPT som används frekvent har en

förmåga att skapa falska referenser gentemot användaren. Anledningen till detta är troligen att chatbotten ska med dess intelligens skapa referenser på egen hand. Detta medför i vissa fall att botten levererar felaktiga referenser på ett självsäkert sätt. Vidare förklarar Olle att det är en risk att AI inom diverse språkmodeller kan medföra att studenter är mer benägna att fuska. Han menar att språkmodeller som Chat-GPT kan användas för att skriva uppsatser på egen hand. Erbjuds inte en utbildning anpassat till den stundande utvecklingen tror Olle att högskolor och universitet kommer att erbjuda en föråldrad utbildning i framtiden. Olle menar att dessa språkmodeller bör användas för lärande istället för fusk. Josefin förklarar att en hög transparens krävs för att uppnå en kontroll kring AI-system men även att ett government system krävs för fullständig kontroll. Vidare i intervjun med Josefin förklarar hon att användare men framförallt utvecklare inte naivt får acceptera det resultat systemen genererar utan bör med ett kritiskt tillvägagångssätt ständigt utveckla systemen så att resultatet blir så optimalt som möjligt. Gällande användare anser Josefin att samtliga resultat från exempelvis Chat-GPT bör kontrolleras för att undvika felaktig eller missvisande information. Josefin tar även upp risken kring att personlig data kan exponeras genom kraftfulla AI-system. För att motverka detta anser Josefin att ett etiskt filter krävs för att säkerställa att outputten är acceptabel. Mathias anser att en funktion som skulle hjälpt språkmodellerna att leverera mer tillfredsställande svar är om en typ av relevansscore var tillgänglig för varje output från AI. Han menar då att vi människor kan bedöma svaren och om resultatet inte ses som acceptabelt eller om det kan skrivas bättre ger man modellerna feedback för att utvecklas.

Författarna Floridi et al (2018) förklarar att en överanvändning av AI och ett naivt tankesätt kan leda till att dessa kraftfulla och superintelligenta system används med fel avsikter. Bedrägerier och manipulation ses som tänkbara effekter om verktyget hamnar i fel händer, samt ifall samhällets synsätt kring AI är naivt och okontrollerat (Floridi et al. 2018). AI är bra på att utföra sina mål men målen behöver inte vara i linje med de mänskliga målen vilket kan leda till katastrofala konsekvenser (Sotala & Yampolskiy, 2014). Josefin förklarar att en risk med AI är ifall det hamnar i fel händer. Hon menar att dessa kraftfulla verktyg är menade för bra ändamål men kan missbrukas av olika aktörer. Däremot anser hon att detta inte är unikt för AI utan gäller samtliga typer av IT-lösningar. En risk ligger även inom möjligheten att skapa egenkomponerade system, eftersom dessa används av individer utan förståelse för hur modellerna i själva verket fungerar kan resultatet oavsett avsikt resultera i felaktig output. Risken kring att AI hamnar i fel händer ser Josefin som något problematiskt. När länder utan liknande öppenhet som Sverige använder AI finns det en risk till spridning av missvisande och skadlig information. Olle menar istället på att fördomar kring att länder som exempelvis Kina inte tar ansvar att utveckla säker och pålitlig AI är felaktig och att länder och ledande AI-företag bör samarbeta för att nå en tillfredsställande nivå.

Kelley et al (2021) undersöker människors tankar kring AI och resultatet visar att 12,2 procent känner en användbarhet i AI och tror starkt på att teknologin kommer vara hjälpsam och assistera människor i deras sätt att utföra uppgifter (Kelley et al. 2021). Li (2021) anser däremot att en användning av AI kommer resultera i att människors utveckling kommer att hämmas. Sociala och etiska egenskaper som människan besitter bör inte utvecklas i enlighet med AI-system eftersom det skulle försämra utveckling (Li, 2021). Josefin anser att AI bör användas som ett verktyg för att assistera människor. Det är ett kraftfullt verktyg som kan delge en mängd information och då krävs det att människor fortsätter att använda sin egna förmåga för att använda detta verktyg på ett korrekt sätt. Hon menar att ett verktyg som Chat-GPT kan ge nya infallsvinklar och idéer men att det kräver att användaren förstår dessa idéer och kan motivera och försvara dem. Verktyget kan ge inspiration för att se nya perspektiv men risken finns att människor slutar eller glömmer att tänka själv. Mathias

instämmer med Josefins tankar och ser verktyget som innovationsdödande om det används på fel sätt. Han menar att kreativiteten, framförallt för yngre människor, kommer bli lidande om man slutar tänka själv och lägger sitt fulla förtroende för verktyg som Chat-GPT.

Det finns ett flertal fall där AI inte har levt upp till förväntningarna och ett av dessa är när Microsoft lanserade chatbotten Tay 2016. Lanseringen gick inte som tänkt och verktyget lades ner efter bara 16 timmar. Anledningen var eftersom Tay började agera hatiskt mot användare som ville kommunicera med botten (Zemčík, 2020). I nutid visar ungefär 85% av samtliga AI lösningar ett partiskt resultat. Det finns även exempel där AI-system har utdaterat data vilket har lett till diskriminering och integritetskränkningar (Brand Studio & anch.AI, 2023). Frågeställningen kring om AI kan ersätta olika typer av arbeten är återigen aktuell där verktyget kan ses som ett hot mot arbetsmarknaden. I intervjun med Josefin förklarar hon att AI är ett verktyg som har oändliga möjligheter och kommer föra samhället framåt. Hon jämför utvecklingen med tidigare revolutionära händelser där samhället utvecklats och nya jobb tillkommit. Josefin anser att AI kommer att ersätta ett flertal jobb eller i vissa fall ändra arbetsuppgifterna i framtiden på samma sätt som revolutioner historiskt har gjort. Historiskt har däremot nya arbeten skapats vid innovation och det tror Josefin kommer bli fallet med AI revolutionen. Att arbetslösheten kommer öka eftersom AI tar över arbetsuppgifter håller inte Josefin med om utan ser utvecklingen som positiv och beroende av mänsklig interaktion. Olle har liknande tankar kring arbetsmarknaden som Josefin eftersom han anser att AI kommer förändra arbetsmarknaden, men även skapa nya arbeten. Olle ser däremot en risk för människor utan tillräcklig utbildning eller omställningsmöjlighet. Han anser att individen behöver anpassa sig till den nya arbetsmarknaden för att utföra framtidens arbetsuppgifter. Mathias beskriver en annan risk, att företag eller individer inom företag för hastigt beslutar sig att använda sig av AI utan att förstå konsekvenserna. Han menar att det är smidigt att använda AI-verktyg men att data som matas in kan vara privat eller konfidentiell. Ger man företagen bakom AI-verktygen tillgång till datan äger de informationen, de kan alltså använda den utan konsekvenser. Utöver denna risk håller Mathias med Olle och Josefin om att jobb kommer försvinna men att nya jobb kommer att skapas. Mattias förklarar att AI-utvecklingen kan jämföras med den digitala revolutionen. Då trodde man att datorerna skulle ta över människors arbete vilket visade sig vara felaktigt. Han menar på att AI kan ersätta arbetsuppgifter samtidigt som nya kommer att skapas men att människan kommer förbli sysselsatt.

AI har potentialen att bidra med revolutionära upptäckter till världen inom exempelvis sjukvård, transport och energiförsörjning. AI-systemen kan även hjälpa företag att fatta bättre beslut med dess intelligens. En förhoppning är även att dessa system kommer lösa problem som fattigdom och klimatpåverkan runt om i världen (European Parliament, 2020). Genom väletablerad design och en mänsklig kontroll kan beslutsmekanismerna resultera i att utdatan inte är felaktig (Sartori & Theodorou, 2022). Värdet av denna teknologi har uppfattats av näringslivet vilket har resulterat i att ett flertal konkurrenter tävlar mot varandra. Denna konkurrenssituation har bidragit till att forskningen, innovationen och utvecklingen har ökat inom området (Smuha, 2021). I det öppna brevet förklaras utvecklingen som utom kontroll eftersom ett flertal aktörer på marknaden tävlar gentemot varandra för att uppnå AGI (Future of Life, 2023). Utvecklingens hastiga fart bidrar med ett par allvarliga konsekvenser som redan visat sig i form av rasistiska chatbottar eller olyckor med självkörande bilar. Om AI närmar sig AGI som har möjlighet att utföra mänskligt arbete bättre än människan ses utvecklingen som ett hot mot mänskligheten (Sotala & Yampolskiy, 2014). Ett citat från AI-forskaren Eliezer Yudkowsky (2008) beskriver AI på följande sätt. “The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something

else.”. Russell[3] (2023) förklarar att AI inte alltid tar hänsyn till de konsekvenser ett genomfört arbete kan medföra. Om AI får en uppgift att lösa ett problem på snabbaste och bästa möjliga sätt utan att ställa motfrågor kan resultatet leda till oönskade konsekvenser (Sweden, S.T.A., Stockholm, 2023). Det krävs alltså någon form av mänsklig förståelse och kontroll för att säkerställa att dessa system inte genomför uppdrag vilket kan leda till olyckliga situationer. En dag kan mänskligheten tappa kontrollen över AI-system via uppkomsten av superintelligenser som inte agerar i enlighet med mänskliga önskemål vilket kan hota vår existens (Russell, Dewey & Tegmark, 2015). Olle förklarar att den konkurrenssituation som skapats mellan AI-företagen leder till att samtliga konsekvenser med utvecklingen inte beaktas ordentligt. Han anser att det primära för dessa företag är att uppnå sitt mål att nå AGI vilket skapar en konsekvens att AI löper större risk att ge manipulerande eller missvisande resultat. Risken kring att dessa aktörer på marknaden inte tog konsekvenser på allvar visade sig när Microsoft lanserade en chatbot som diskriminerade användare. Olle tycker att denna typ av AI inte är mer farlig än andra typer av sökningar på internet. Han menar däremot att det tankesättet att nå utveckling och framgång till varje pris kommer bli ett allvarligt problem när mer kraftfulla typer av AI lanseras. Den största risken Olle ser med utvecklingen är att det kan utvecklas en AI som kommer förinta mänskligheten. När det är garanterat att AI inte kommer förinta mänskligheten anser Olle att vi bör fortsätta utvecklingen. Han anser att utevecklingspausen är en nödvändighet och får ta den tid som krävs för att stabilisera utvecklingen igen. Olle ser däremot utvecklingens möjligheter som en räddare för framtiden om utvecklingen genomförs på ett kontrollerat sätt. Varken Josefin, Mathias eller Mattias delar samma syn som Olle om att det finns en risk att AI kan förinta mänskligheten utan anser att vi bör vara glada över den utvecklingen som kan leda till ofantliga möjligheter. Mathias har en liknande syn som Josefin, han förklarar även att de som skrivit under det AI-upproret är kompetenta människor inom sina områden och att de kan se något han inte ser.

[3] Stuart Russell, Intervju den 7 Februari 2023, https://www.svtplay.se/video/jNnWaXy/anders-hansen-moter/stuart-russell

7. Slutsatser

Syftet med denna uppsats har varit att undersöka osäkerheten kring AI-utvecklingen och de risker som kan uppkomma om AI fortsätter att utvecklas utan tillräcklig mänsklig förståelse. För att besvara dessa forskningsfrågor har det genomförts en kvalitativ studie där experter inom ämnet intervjuats och relevanta artiklar och expertartiklar har analyserats.

I detta kapitel kommer resultaten att sammanfattas och slutsatser kommer att dras utifrån analysen. Dessutom kommer kapitlets struktur att beskrivas och hur de två forskningsfrågorna har besvarats. Slutligen presenteras en potentiell modell för att hantera riskerna och osäkerheten med AI som kan ligga till grund för framtida forskning.

7.1 Slutsatser utifrån forskningsfrågorna

Är osäkerheten med AI-utvecklingen befogad?

Kapplöpningen som skapats av de stora AI-bolagen har utan tvekan bidragit till den snabba utvecklingen och ökat osäkerheten kring vad AI kommer leda till. Future of Life (2023) påpekar denna osäkerhet där många experter inom området delar samma syn. Samtidigt som utvecklingen rasar fram ser fortfarande andra experter detta som en utveckling som inte går att hindra och försöka pausa eller stoppa den kan leda till missbruk av tekniken. Det lyder inga tvivel om att det finns en osäkerhet kring utvecklingen av AI enligt denna undersökning. Hur utvecklingen kommer att påverka samhället finns det många åsikter om, från människans undergång till en teknik som kommer att lösa alla världens problem. Flera forskare inom ämnet tror att den rådande uppmärksamhet AI har fått inom media kan skapa större osäkerhet kring AI samtidigt som det kan bidra med att ge folk mer försiktighet vid användandet. Enligt studien visar det sig att forskare tror starkt på att en reglering om hur AI-utveckling bör gå till så att tekniken arbetar i linje med människans välfärd och mål. Forskare poängterar att det skulle krävas en reglering på världsomspännande nivå för att en reglering inte ska ses som onödig eller farlig. Det öppna brevet Future of Life (2023) publicerat har skrivits under av många experter och teknikintresserade individer vilket tyder på att det finns en osäkerhet kring utvecklingen. Alla undertecknare av brevet är inte eniga om det framlagda förslaget men de tycker att det är ett steg i rätt riktning. Enligt denna studiens undersökning visar resultatet att en utvecklingspaus inte är nödvändig i dagsläget. Det som krävs för att minska osäkerheten är att öka förståelsen, transparensen och uppmärksamma riskerna kring AI.

Forskare kommer i framtiden kunna utveckla autonoma AI-lösningar och det kommer finnas en möjlighet till att helautomatisera samhället med hjälp av AI. Resultatet från denna undersökning visar trots att AI kommer kunna fungera autonomt i framtiden menar respondenterna att detta inte är fallet. Det kommer alltid vara nödvändigt med en mänsklig interaktion tillsammans med teknologin då det alltid kommer finnas en osäkerhet med att systemen kan göra fel. AI kommer heller aldrig kunna utvecklas på ett sätt så att den kommer kunna tänka och bete sig exakt som en människa och därför kommer det alltid behövas en övervakning av systemen. Det kan även hända impulsiva och drastiska händelser i verkligheten som AI inte kan ta hänsyn till och därför kommer det alltid vara nödvändigt med en symbios mellan teknologin och människan.

Vilka risker kan uppkomma med AI:s fortsatta okontrollerade utveckling utan mänsklig förståelse?

Informationen som samlats in under litteraturdelen och de genomförda intervjuerna fastslår ett resultat som visar att AI kan under den hastiga utvecklingen generera kränkande, felaktig eller missvisande information. Resultatet visar att AI inte alltid har förmågan att skilja på säker respektive osäker data vilket också kan medföra att AI genererar kränkande, felaktig eller missvisande information. Resultat från AI kan även vara missvisande eller direkt felaktiga. Oavsett om det bristfälliga resultatet är en effekt av hastig utveckling och ett undermåligt konsekvenstänk eller om AI-systemen använder sig av utdaterad data leder detta till att AI kan leverera felaktiga eller missvisande svar. Det framförs även att kraftfulla system har möjligheten att kränka individens privata information. Resultatet visar att användning av utdaterad data inom AI-system kan leda till integritetskränkningar som grundar sig på ålderstigna normer kring exempelvis etnicitet och kön. En hög transparens krävs för att säkerställa att modellerna inte agerar på detta oönskade sätt.

AI-systemen är ett potentiellt hot mot utbildning då språkmodeller som Chat-GPT kan användas för att fuska. Dessa modeller har möjligheten att skriva klart uppsatser utan att användare själva behöver komma med egna tankar eller idéer. Trots att detta ses som ett hot visar resultatet att verktyget kan användas på ett sätt som istället omvandlar hotet till en möjlighet. Dessa system kan assistera individen om de används på rätt sätt. Om inte utbildningssektorn anpassar sig mot utvecklingen kommer en utdaterad utbildning erbjudas. Risken att studenter eller yngre människor tappar kreativitet och innovationstänkande eftersom det tillförlitar sig på dessa språkmodeller är hög. Det är smidigt att använda ett verktyg utan att behöva ifrågasätta svar eller själv behöva motivera sina tankar eller ideer men resultatet visar att det finns en risk att detta kan leda till att dagens yngre användare blir liktänkande och kommer sakna en källkritisk syn.

Resultatet visar att risker som bedrägerier eller manipulation kan ske om dessa kraftfulla verktyg hamnar i fel händer. Verktyget får inte användas naivt utan ett konsekvenstänk bör alltid vara centralt. Resultatet visar även att fördelaktig innovation alltid kan användas på fel sätt oavsett om de missbrukas avsiktligt eller inte av diverse aktörer. Det är en oenighet kring hur vi kan lita på att länder som Kina eller Ryssland utvecklar dessa system dels med en öppenhet men även med en positiv avsikt. Ett samarbete mellan samtliga aktörer ses som en lösning på detta problem, dels för att se över utvecklingen och säkerställa att samtliga länder inte kliver utanför en bestämd acceptabel nivå. Detta för att minska risken att AI utvecklas med onda intentioner.

Studiens resultat visar att det finns en risk att AI-systemens utveckling kommer att ersätta ett antal arbeten och arbetsuppgifter eftersom AI kommer konkurrera ut människan med dess effektivitet. Detta ses som naturligt och trots att risken är befogad förklaras risken som en del av utvecklingen. Risken kring att företag eller individer inte anpassar sig till utvecklingen är också befogad vilket kan ställa till stora problem. Företag kommer behöva anpassa sig till utvecklingen för att fortsätta vara relevanta annars finns risken till att företag kommer behöva läggas ner. Detta gäller även individen, resultatet visar att förståelsen kring AI och högre utbildningskrav kommer att krävas i framtiden vilket kommer göra det mer komplicerat att få ett arbete inom dessa områden.

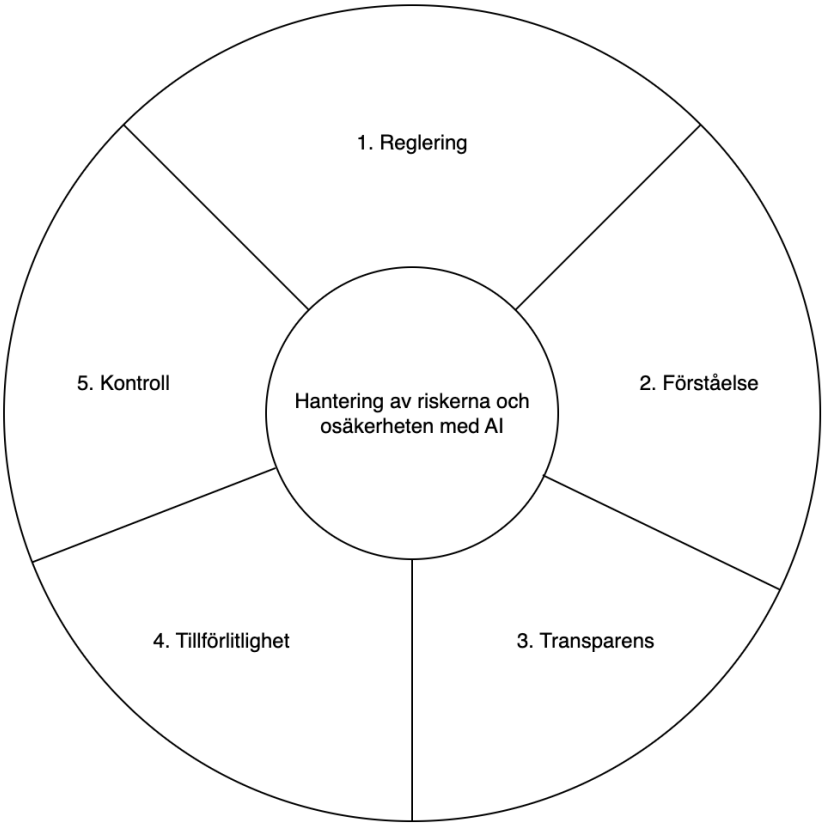
Den värsta tänkbara risken som undersökningen visar är möjligheten att AI och dess superintelligens i framtiden kommer konkurrera ut människan och förinta mänskligheten.

Resultatet visar att denna risk uppstår eftersom ett antal företag tävlar mot varandra för att leverera bästa möjliga produkt på marknaden och strävar efter att bli marknadsledande. Kapplöpningen medför att företag lägger mindre fokus på säkerhet eller transparens och mer på att nå så hög teknisk utveckling som möjligt. För att minska denna risk behöver utvecklingen ske med ett kritiskt synsätt. Utvecklas inte dessa avancerade system korrekt utan att det fortsätter med ett långsiktigt tankesätt att utveckla tekniken till varje pris ökar risken att mänskligheten i framtiden kommer förlora kontrollen. Om utvecklingen sker okontrollerat finns det också en risk att teknologin blir ett hot mot mänskligheten istället för den tänkta intentionen. Resultatet visar dock på att människor inte behöver vara oroliga för att AI i framtiden kommer att ses som ett hot mot mänskligheten. Däremot anses teknologin vara direkt farlig om tillräckligt avancerade system används med fel intentioner, det är alltså inte AI vi ska vara oroliga för utan människor som kan tänkas använda dem.

7.2 Framtida forskning

AI är ett område som har stora framtida forskningsmöjligheter då ämnet är brett och ständigt utvecklas. Denna studiens syfte har varit att undersöka om osäkerheten kring AI-utvecklingen är befogad samt vilka risker som kan uppkomma med utvecklingen. Från studiens resultat, analys, diskussion och slutsatser har en femstegsmodell utvecklats som förslag för hur man kan hantera osäkerheten och riskerna med AI och den stundande utvecklingen, se Figur 1. Modellen har inte tillämpats eller undersökts i denna studie då det legat utanför studiens ramar men är ett intressant framtida forskningsområde. Femstegsmodellen kan användas för framtida forskning genom att tillämpa modellen och analysera hur faktorerna reglering, förståelse, transparens, tillförlitlighet och kontroll kan hjälpa till att hantera osäkerheten och riskerna kring AI. Dessa faktorer har tagits fram efter att analyserat vanligt förekommande fenomen från den insamlade litteraturen samt intervjuerna där AI har en tydlig påverkan på samhället.

Figur 1 - Femstegsmodell



Då denna studien är avgränsad till att endast undersöka experters kunskap inom området är ett annat intressant förslag till framtida forskning att undersöka hur osäkerheten och riskerna kring AI uppfattas generellt i samhället. Ett annat intressant förslag är att utföra samma undersökning fast på andra geografiska platser för att se och kunna jämföra hur osäkerheten och riskerna skiljer sig runtom i världen. Framtida forskning inom dessa områden möjliggör en bredare och djupare förståelse för forskningsområdet.

8. Referenser

Bengt Starrin & Per-Gunnar Svensson (2008). *Kvalitativ metod och vetenskapsteori.* Lund: Studentlitteratur.

Brand Studio & anch.AI (2023). Människan ska styra AI – inte tvärtom https://www.di.se/brandstudio/anch-ai/manniskan-ska-styra-ai-inte-tvartom/ [Hämtad: 2023-04-21]

Bubeck, Sébastien. et al. (2023) Sparks of Artificial General Intelligence: Early experiments with GPT-4. Microsoft Research. [online]. doi: 10.48550/arxiv.2303.12712

Conn, A. (2015) Benefits & Risks of Artificial Intelligence, *Future of Life Institute,* 14 november. https://futureoflife.org/ai/benefits-risks-of-artificial-intelligence/ [Hämtad: 2023-04-21]

Du-Harpur, X., Watt, F.M., Luscombe, N.M. & Lynch, M.D. (2020) What is AI? Applications of artificial intelligence to dermatology. *British journal of dermatology (1951)*. [Online] 183 (3), 423–430. doi:https://doi-org.lib.costello.pub.hb.se/10.1111/bjd.18880

European Parliament, Directorate-General for Internal Policies of the Union, Eager, J., Whittle, M., Smit, J., et al. (2020) Opportunities of artificial intelligence. European Parliament. doi:https://data.europa.eu/doi/10.2861/692222

Fietta, V., Zecchinato, F., Stasi, B., Polato, M. & Monaro, M. (2022) Dissociation Between Users' Explicit and Implicit Attitudes Toward Artificial Intelligence: An Experimental Study. IEEE transactions on human-machine systems. [Online] 52 (3), 481–489. doi:10.1109/THMS.2021.3125280

Floridi, L., Cowls, J., Beltrametti, M. *et al.* AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds & Machines* 28, 689–707 (2018). doi:https://doi.org/10.1007/s11023-018-9482-5

Future of Life. (2023). Pause giant AI experiments: An open letter. 22 mars. https://futureoflife.org/open-letter/pause-giant-ai-experiments/ [Hämtad: 2023-05-12].

Fröding, B. and Peterson, M. (2020). Friendly AI. Ethics and Information Technology. doi:https://doi.org/10.1007/s10676-020-09556-w.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines, 30, pp.99–120. doi:https://doi.org/10.1007/s11023-020-09517-8.

Hsieh, H.-F. & Shannon, S. E. (2005) Three Approaches to Qualitative Content Analysis. Qualitative health research. [Online] 15 (9), 1277–1288. doi:https://doi.org/10.1177/1049732305276687

Jacobsen, D. I. & Andersson, S. (2017) Hur genomför man undersökningar? : introduktion till samhällsvetenskapliga metoder. 2 uppl. Lund: Studentlitteratur AB.

Kelley, P. G., Yang, Y., Heldreth, C., Moessner, C., Sedley, A., Kramm, A., Newman, D. T. & Woodruff, A. (2021) 'Exciting, Useful, Worrying, Futuristic: Public Perception of Artificial Intelligence in 8 Countries', in AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. [Online]. 2021 Ithaca: Cornell University Library, arXiv.org. pp. 627–637.

Li, O. (2021). Problems with 'Friendly AI'. Ethics and Information Technology. doi:https://doi.org/10.1007/s10676-021-09595-x.

Lincoln, Y. S. & Guba, E. G. (1985) Naturalistic inquiry. Beverly Hills, Calif: Sage.

Liu, B. (2021) In AI We Trust? Effects of Agency Locus and Transparency on Uncertainty Reduction in Human–AI Interaction. Journal of computer-mediated communication. [Online] 26 (6), 384–402. doi:10.1093/jcmc/zmab013

Mirbabaie, M., Brünker, F., Möllmann Frick, N.R.J. & Stieglits, S. (2022) The rise of artificial intelligence – understanding the AI identity threat at the workplace. Electronic markets. [Online] 32 (1), 73–99. doi:10.1007/s12525-021-00496-x.

Morales-Forero, A., Basetto, S. & Coatanea, E. (2023) Toward safe AI. *AI & society*. [Online] 38 (2), 685–696.

Morikawa, M. (2017) Firms expectations about the impact of ai and robotics: Evidence from a survey. *Economic inquiry*. [Online] 55 (2), 1054–1063. doi: https://doi-org.lib.costello.pub.hb.se/10.1111/ecin.12412

OpenAI (2023). GPT-4 Technical Report. arXiv (Cornell University). doi:https://doi.org/10.48550/arxiv.2303.08774.

Patel, R. & Davidson, B. (2019). Forskningsmetodikens grunder: att planera, genomföra och rapportera en undersökning. Lund: Studentlitteratur

Russell, S., Dewey, D. and Tegmark, M. (2015). Research Priorities for Robust and Beneficial Artificial Intelligence. AI Magazine, [online] 36(4), p.105. doi:https://doi.org/10.1609/aimag.v36i4.2577.

Russell, S. (2021). The history and future of AI
Oxford Review Of Economic Policy. doi:https://doi.org/10.1093/oxrep/grab013.

Ryan, M. (2020) In AI We Trust : Ethics, Artificial Intelligence, and Reliability. *Science and engineering ethics*. [Online] 26 (5), 2749–2767. doi:10.1007/s11948-020-00228-y

Sartori, L. and Theodorou, A. (2022). A sociotechnical perspective for the future of AI: narratives, inequalities, and human control. Ethics and Information Technology, 24(1). doi:https://doi.org/10.1007/s10676-022-09624-3.

Smuha, N. A. (2021) From a 'race to AI' to a 'race to AI regulation': regulatory competition for artificial intelligence. Law, innovation and technology. [Online] 13 (1), 57–84. doi:https://dx.doi.org/10.2139/ssrn.3501410

Sweden, S.T.A., Stockholm (2023). Anders Hansen möter... – Stuart Russell. [online] www.svtplay.se. Available at: https://www.svtplay.se/video/jNnWaXy/anders-hansen-moter/stuart-russell [Hämtad: 2023-04-27]

Tamboli, A. (2019) *Keeping Your AI Under Control A Pragmatic Guide to Identifying, Evaluating, and Quantifying Risks*. 1st ed. 2019. [Online]. Berkeley, CA: Apress.

Techworld (2021). *Det här är AI och så funkar det.* https://techworld.idg.se/2.2524/1.699032/ai-sa-funkar [Hämtad: 2023-04-27]

Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G. & Kaplan, L. (2020) Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI. Patterns (New York, N.Y.). [Online] 1 (4), 100049–100049. doi:10.1016/j.patter.2020.100049.

Wu, J. & Shang, S. (2020) Managing uncertainty in ai-enabled decision making and achieving sustainability. Sustainability (Basel, Switzerland). [Online] 12 (21), 1–17. doi:https://doi.org/10.3390/su12218758

Yampolskiy, R. V. (2019) Predicting future AI failures from historic examples. Foresight (Cambridge). [Online] 21 (1), 138–152. doi:10.1108/FS-04-2018-0034

Yudkowsky, E. (2008) Artificial intelligence as a Positive and Negative Factor in Global Risk. Global Catastrophic Risks, 308-345.

Zemčík, T. (2020). Failure of chatbot Tay was evil, ugliness and uselessness in its nature or do we judge it through cognitive shortcuts and biases? AI & SOCIETY. doi:https://doi.org/10.1007/s00146-020-01053-4.

9. Bilagor

9.1 Bilaga 1: Transkribering från intervjun med Olle Häggström

Ludwig: Vi har egentligen 3 områden vi kommer gå in på, osäkerheten kring AI utvecklingen, sen risker/problematik och sist konsekvenser. Det vi vill börja med är att fråga om dina tankar om själva osäkerheten kring AI-utvecklingen? Är det något du anser vi behöver vara oroliga för?

Olle: Svaret på den sista frågan är ja, vi behöver vara oroliga. Formulera början på första frågan en gång till.

Ludwig: Vi tänker just dina tankar kring osäkerheten med AI-utvecklingen, det som har blivit väldigt aktuellt den senaste tiden.

Olle: En central del av problematiken det är ju det här med artificiell generell intelligens och om och när vi kan väntas nå ett sånt genombrott. Och jag har jobbat med dem här AI-frågorna i över ett decennium och då har frågan om AGI jag har alltid haft det uppe på min radar för att det varit väldigt intressant men det verkat som den legat decennier fram i tiden asså minst 20-30 år, kanske längre 50 år eller 100 år eller mer. Men dem senaste åren, ungefär pandemi åren och framåt så har utvecklingen accelererat på ett sånt sätt så att för det första det första så tycker jag inte att det längre finns någon rimlig tvivel om att AGI är möjligt. Det är liksom att vi är på väg mot det och för det andra att vi kan vara på väg mot ett mycket snabbare än vad vi tidigare har tänkt oss. Så att när om ni frågar mig tidigast när AGI kan komma så har jag inget annat svar än på det än imorgon. Jag tror ju inte att det kommer imorgon men den kan göra det, den kan komma om 1 år och den kan komma om 5 år eller om 10 år. Och sen så är det väl nästan det troligaste att det kommer inom det närmaste decenniet och min syn på det här med AGI då är att när vi skapar det då är det det största paradigmskiftet någonsin i vetenskapens historia. Därför att det är ändå så att det här med hjälp av vår intelligens som vi har skaffat oss vår unikt starka position, vi dominerar den här planeten. Och när vi då automatiserar denna otroliga resurs och hamnar i ett läge där vi går ner till andra plats i rangordningen om varelsen med högsta intelligensen då är det ett så genomgripande steg så att industraliseringen eller jordbruksrevolutionen det kan slänga sig i väggen. Och det kan bli fantastiskt bra, det kan bli lösningen på alla våra problem och det kan bli slutet för mänskligheten om vi gör det på fel sätt. ‘

Ludwig: Men skulle du anse att AI-utvecklingen är utom kontroll?

Olle: Säkert är det absolut inte, nej, nej, nej, nej. Utom kontroll jag menar det beror på vad vi gör, jag vägrar släppa tanken om att vi kan ta oss samman och börja agera mer ansvarsfullt än vad vi gjort hittills. Men tittar man på hur dem ledande AI-företagen agerar idag så är det inte på den banan vi är nu riktigt och ett av syftena med brevet är att ta upp diskussionen och hjälpa oss att komma på banan då. Jag tror inte man ska hänga upp sig för mycket på det

liksom konkreta förslaget om en 6 månaders paus för träning av stora modeller, det är egentligen inte det som brevet handlar om, det handlar mer om att lyfta upp till bredare diskussioner hur oacceptabelt stora dessa riskerna är.

Ludwig: Men inom vilka områden känner du inom AI det finns störst chans till att uppleva denna osäkerheten vi pratar om?

Olle: Jag tror så här, det är dem stora språkmodellerna och relaterat generativ AI som ligger i framkant nu och som är den troligaste vägen fram till AGI. Och det är där liksom den allra största osäkerheten ligger. Vad kommer det här leda till?

Ludwig: Sen har vi en annan fråga, tror du vi kommer kunna leva i ett samhälle där AI får leva sitt egna liv eller kommer det alltid behövas en "human in the loop"?

Olle: Det är en bra fråga, iprincip så tror jag att det är möjligt att designa ett utopiskt samhälle där AI gör allting åt oss, där vi inte behöver vara med och fatta beslut, därför att AI gör det åt oss och vi har sett till att AI hela tiden har vårt bästa för att göra det åt oss, men det känns inte som att det är det idealiska samhället för det känns som att mänskligt liv i ett sånt samhälle att det saknas någonting, agens, mening så att vi kommer antagligen att vilja vara med i det här beslutsfattandet.

Marcus: Det blir som att man utvecklar AI till att funka mot en funktion hela tiden ändå asså att den inte är wild eller bildlig i samhället eller så, att den fyller en funktion för de flesta om man säger så.

Olle: Jag tror att när vi väl har en superintelligent AGI då kommer den vara wild i den meningen att den kan göra vad den vill. Men en tanke skulle kunna vara då att om den har ändå viljan att släppa in human in the loop i vissa slags frågor så kan den göra det. Men ett annat typ av scenario är faktiskt att AI släpps loss helt och hållet, vi skiter i det här med human in the loop därför att det finns något lite konstigt med att att insistera på human in the loop i ett läge där Ai är en bättre beslutsfattare i alla frågor. Bättre än vi på att tänka ut ett svar då blir det som att vi vill gärna insistera på att vara med i loopen bara till för ett skadligt brus och därför skulle man kunna föreställa sig en värld där vi får utlopp för vårt behov av agens, och beslutsfattande på liksom särskilda avskärmade arenor lite grann med analogt med om vi tänker oss utvecklingen av autonoma fordon så kan man tänka sig en framtid, som stor sannolikt då att vi om 15-20 år inte överhuvudtaget tillåta mänskliga bilförare på allmän väg därför för att det är för trafikfarligt när vi har autonoma fordon som kör bättre än vi, då vill vi inte utsätta allmänheten för den risken, extra risk som det är med mänskliga bilförare men folk kommer ändå vilja köra bil och då kanske man ska få göra det men det kommer isåfall vara på särskilda inhägnade banor då.

Olle: På samma sätt skulle man titta på generellt mänskligt beslutsfattande som skulle kunna ske inom säkra... Dataspel helt enkelt. Avancerade dataspel, avancerade virtuella världar, där

vi kan vara alla möjliga äventyr, där det krävs, där våra sinnen ställs på prov i att fatta svåra beslut, men ingenting är egentligen farligt. Så skulle framtiden kunna se ut.

Ludwig: Men vad tror du hade krävts av oss människor när det kommer till utvecklingen av AI, för att minska den här osäkerheten. Finns det något man kan göra framtidsmässigt?

Olle: Asså jag, när du säger minska osäkerheten så tolkar jag det som att ta bort dem värsta riskerna.

Ludwig: Precis

Olle: Det är det viktigaste.

Ludwig: och när du pratar om dem här riskerna, vilka risker pratar du om då?

Olle: Det finns alla möjliga typer av risker. Alla risker som handlar om att: AI tar beslut som är diskriminerande eller AI.. eller det senaste som börjat diskuterats är att, folket i 3:e världen anställs under svåra förhållanden och ger human feedback till reinforcement learning träningsprocesser och sånna här saker. och det finns frågor kring: kommer AI skapa / användas för dessinformation. Alla dem frågorna, de är viktiga men, frågan är hurvida AI kommer förinta mänskligheten är större och viktigare och den är en överhängande. För att om vi inte lyckas lösa den frågan om mänskligheten överlever så blir alla de andra frågorna.. ah.. de finns ju inte längre. Så om jag får lov att fokusera på en fråga i det här samtalet: “Hur ser vi till att AI inte dödar samtliga människor”. Och det är default utfallet om vi fortsätter som nu dvs. om vi inte tar..
Behärskar ni termen AI-alignment?

Marcus: till en liten grad. Det är väl att den ska följa.. eller att den ska hjälpa oss mer än den ska.. ah den ska följa människans riktlinjer.

Olle: Ja vi ska hitta ett sätt att se till att de första superintelligenta maskinerna har mål och drivkrafter som i tillräcklig mån prioriterar mänsklig välfärd och, vilka de nu är, mänskliga värderingar som vi vill att maskinerna ska ha. Det är att se till att det blir så. Det är det som är forskningsområden i AI-alignment. Om vi inte lyckas med detta, om vi struntar i det, som är vad vi i stort sett gör nu. Det jobbar tiotusentals eller hundratusentals människor med AI utveckling idag. Bara några hundra som är inom AI-alignment. Det är liksom galna proportioner. Om vi fortsätter att strunta i detta så är default utfallet, det som Eliezer Yudkowsky formulerade med orden: AI hatar dig inte, den älskar dig inte heller, men du är gjord av atomer som den kan ha annan användning för. Olika varianter av det är vad som troligen kommer att hända om vi inte tar oss samman och löser det här med AI-alignment, innan vi kommer fram till AGI. Det finns två olika vägar eller en kombination utav dem för att nå fram till det här att vi inte blir dödade av AI. Den ena är att vi löser AI-alignment, det är ett svårt, svårt teknisk problematik. De flesta verkliga resultat i de här sakta växande eller ganska snabbt växande procentuellt sett fast det är fortfarande litet, forskningsområdet. De

flesta resultat pekar riktning mot hinder på vägen, teoretiska svårigheter med att faktiskt lösa det här. Så vi är långt ifrån att se en tydlig väg framåt för att lösa AI-alignment. Men det är ändå hoppet att fixa det. Det är den ena lösningen. Den andra lösningen är att inte bygga AGI, det är också jätte, jätte svårt. Därför det kräver social och politisk koordination kring den uppgifter. Därför att, hur det ser ut nu, befinner sig dem stora AI bolagen i en kapplöpning med varandra. Det gäller att roffa åt sig marknadsandelar och marknadsdominans och det här pressar dem att köra framåt. De vill inte bli sinkade av etiska överväganden och AI-alignment problematik och sånt utan de vill köra rakt fram.

Ludwig: Det kan också vara en stor risk.

Olle: Ja, det är en katastrofal risk. Och det finns kapplöpning på flera plan också. Det finns de som kritiserar det här FLI brevet: "jaja om vi i väst lyckas göra det här, så finns ju kina och andra kanske kör om oss då" Och då vill jag understryka att det finns ingenting i FLI-brevet som säger att "Vi vill att amerikanska och europeiska AI företag ska göra så här och göra den här 6 månaders pausen". Vi vänder ju oss till alla ledande AI företag. Och jag gillar inte den typen av fördomsfullhet där man tar förgivet att inte kinesiska företag skulle kunna ta ett ansvar här för att undvika mänsklighetens undergång. Så det här är då sammanfattningsvis då: två vägar framåt, lösa AI-alignment eller se till att inte bygga AGI. Sen tror jag i praktik, det jag ör mest optimistisk om, är ju en kombination här, där vi lyckas uppskjuta AGI tillräckligt länge för att vi ska ha hunnit hitta vägar att lösa AI-alignment

Ludwig: Men det är dem två stora faktorerna du känner är viktigast då?

Olle: Ja

Marcus: Då går vi in lite på datan som man kan få fram, speciellt med chat-gpt och de här självgenerande chatbottarna och sånt där. Hur ska man kunna lita på det resultat både över. eller ska man framställa det tydligare att de här kan vara missinformation. Det känns inte som dagens samhälle är så jätte kunniga om själva AI om man säger så och jag vet inte vilken stor säkerhet de känner i användning av sådana här saker. Så man vet inte riktigt hur säkra den informationen om det inte framgår. Du prata lite om att t.ex. källhänvisningar kan vara påhittade eller att de bara liknar att den har kopierat. Vet inte hur jag ska ställa frågan.

Olle: Ja men jag förstår. Opålitlighetfrågan och risken att vi överskattar tillförlitligheten i det vi får ut. Det här är inget nytt problem överhuvudtaget, det har funnits alltid, för att människor är opålitliga. Eller hur?
Det förekommer att människor ger falsk information, det är mycket vanligt till och med. Så i princip skulle jag nog kunna säga, att applicera det sunda förnuft och källkritik och sånt där som vi har gentemot människor, det kan vi applicera på nya AI-modellerna. Men problemet är det är att vi har en extremt snabb teknikutveckling som gör att vi behöver väldigt snabbt uppdatera på hur dem här systemen verkligen fungerar när de är pålitliga och inte pålitliga och så vidare. Så problematiken ställs lite på sin spets av den snabba teknikutvecklingen.

Marcus: Jag tänker lite på det också med maskinlärd AI där AI lär sig själv mycket, det betyder väl också egentligen då att den ska lära sig utifrån vad vi gör så att då måste människan vara perfekt för att AI ska bete sig på ett korrekt sätt.

Olle: Det behöver ju inte vara så, det finns ju fall när man tränar en schack robot tex så kan ju den på egen hand undersöka det offentliga rummet av möjliga schack partier och spelar mot sig själva. De är ju inte alls beroende av mänsklig input. Det var så deep mind byggde den starkaste schackmotorn som finns nudå, alphazero. Med dom principerna att ingen mänsklig input överhuvud taget. Men när det gäller dom stora språk modellerna och så visst, det är mycket mänsklig input där.

Marcus: När vi kollar lite på konsekvenser kring utvecklingen speciellt den här då 6 månaders uppehållet dom skulle föredra. Känner du det är nödvändigt eller tror du AI kan utvecklas som den görs idag utan större konsekvenser?

Olle: Ojojoj, tycker du inte jag har svarat på den frågan?

Marcus: Jo men vi tänker också konsekvenser, det kan vara svårt att förutse konsekvenser också om man inte utvecklar det samtidigt om man säger så?

Olle: Ja det är ett typiskt motargument, och det tycker jag är en väldigt lätt.. Om ni vill fördjupa er i den här frågan så tycker jag ni ska läsa Scott Alexanders kommentar till OpenAIs dokument där han beskriver deras planer framåt då. Dom siktar ju då på att lösa AGI och det är ju explicit deras plan då. Dom går så snabbt framåt de kan och man kan analysera olika argument för det här men jag tycker att argumenten för att bromsa är mycket mycket större än att ta de här argumenten och fortsätta framåt då. De är att vi kommer vara bättre skickade att lösa hela alignment problematiken ju närmare AGI som vi kan experimentera med ligger, svagt argument och ska Scott Alexander på starco… nej astralcodexten heter den. Han diskuterar det där ganska ingående så att jag föreslår att om ni är intresserade av det ska ni läsa den. Nu ska vi se, jag tappade bort mig kring frågan lite grann.

Marcus: Det var väl egentligen hur nödvändigt den här 6 månaders pausen är.

Olle: Juste, det är ju inte exakt nödvändigt att inbromsningen tar exakt den formen vi föreslår, i själva verket ser jag det som osannolikt.

Marcus: Det är la mer att man vill lägga mer vikt på effekterna kring utvecklingen, att man vill undersöka det kanske?

Olle: Jag vill mer säga det som att det behövs en inbromsning. En inbromsning som är tillräcklig för att utvecklarna ska kunna lägga den kraft på säkerställanden av modellernas säkerhet och på AI alignment som verkligen behövs. Jag tror absolut inte att en 6 månaders paus såsom riskiserarna där i det öppna brevet. Att det skulle vara tillräckligt, tror jag absolut

inte, det behövs betydligt mycket mer åtgärder. Jag skrev på brevet ändå därför att jag tycker att det är ett steg i rätt riktning men det behövs mycket mycket mer. Dagen efter brevet så publicerade Yudkowsky en artikel i Time där han förklarade varför han inte skrivit på brevet. Han förklarar även att han välkomnar brevet och att det är ett steg i rätt riktning men det behövs så ofantligt mycket mer. Läs det, för jag är i stort sett på hans linje här. Det behövs ett stopp utav träning av riktigt stora AI modeller som inte har någon bortre tidsgräns och det här stoppet behöver implementeras med kraft med internationella överenskommelser med militär styrka bakom så att de som bryter mot dem vet vad som händer om man bryter mot dem. Därför de som bryter mot dessa överenskommelser sätter hela mänskligheten i fara. Det kan inte tolereras.

Wictor: Men så kort det du säger är att det är ett steg i rätt riktning men att det är ett alldeles för litet steg då?

Olle: Ja, vi behöver mera.

Marcus: Vi har ju gått in på det här med den värsta tänkbara konsekvensen är att den utrotar eller ersätter människan. Så att innan det här AGI blev så stort , eller ja, inte stort så sätt men att det blev en samhällsfråga för allmänheten om man säger så. Jag vet inte om det varit så på tapeten riktigt innan för en vanlig arbetare kanske som inte är insatt i teknologin eller liknande.

Olle: Nej Nej, det här är ju något som hänt under bara de senaste månaderna, tom veckorna att det kommit upp till allmän debatt.

Marcus: Precis och då var vi lite inne på det här att dagens arbetare kanske är lite mer rädd för att AI kommer ta över deras jobb och lite på den banan men då finns det en mycket större konsekvens som kanske inte dom flesta är medvetna om. Om dom inte kollat på någon sci-fi film eller något sånt där, en gammal härlig film. Jag vet inte om vi har en speciell fråga på det men…

Olle: Får jag ställa en fråga till er?

WML: Yes, ja absolut!

Olle: Blir ni, hur reagerar ni på det jag säger?

Marcus: Det är lite skrämmande är det.

Wictor: Ja, det är skrämmande!

Marcus: Men man ser ju också den här utvecklingen som har skett, så man vet ju att den är lite okontrollerbar känns det som just nu.

Olle: Jag skulle hellre säga okontrollerad än okontrollerbar därför att jag inte liksom inte tar för givet att vi inte tar oss samman och faktiskt fixar detta. Men läget är besvärligt.

Marcus: Men man tänker ju väldigt mycket på om det ens är möjligt att uppnå den liksom begränsningen, det känns som att det behöver bli någon begränsning om vad AI ska kunna få utföra. Om det ens är möjligt att nå upp till den nivån liksom. Utan att den liksom ska kunna lägga sina egna regler och nå de värsta konsekvenserna så att säga. Det är ju ett intressant ämne och det är väldigt svårt att greppa det sådär.

Olle: Ja, det kan man ju säga, vi lever i intressanta tider. Det sägs någonstans att det där är något Asiatiskt ordspråk som används, må du leva i intressanta tider. Det är ju lite grann av en förbannelse.

Marcus: Vi hade egentligen inte några mer frågor på det här.

Olle: Nej men va bra. Då hoppas jag att ni hade glädje av detta.

Wictor: Yes absolut det hade vi.

Marcus: Tack så jättemycket för att du ställde upp.

Olle: Ja, ingen fara.

Ludwig: Vi är tacksamma för din tid.

Olle: Ja, okej, tack, hej.

Ludwig: Tack så jättemycket!

9.2 Bilaga 2: Transkribering från intervjun med Josefin Rosen

Ludwig: Så vi tänker lite att vi kan börja med att fråga dig vad dina tankar är om osäkerheten kring AI utvecklingen, den stundade AI utvecklingen.

Josefin: Ja, asså jag tycker ju inte att man varken borde eller kan stoppa i det läge som är nu, absolut inte, snarare det vi ser med stora språkmodeller som Chat GPT och senaste GPT4 och som alla andra, det är försök till liknande är snarare liksom att det är en fördel. Det är en demokratisering av AI med jätte jätte jätte möjligheter för både bolag och privatpersoner att göra bra saker såklart. Men sen är det ju jättestort fokus på vikten av att det behövs reglering, vi har ju ett regelverk som kommer från det så småningom, förhoppningsvis under nästa år och det blir ju viktigare än någonsin med den typen av reglering för allting som är högrisk och sen är det också superviktigt att man sprider kunskapen om hur man använder dessa verktyg på rätt sätt så folk är medvetna om risker och vad man kan lita på och vad man inte kan lita på. Så jag tror det snarare behöver läggas fokus på hur exempelvis studenter hur man behöver anpassa examinationer tex att man behöver beskrivning till kunder om det är så att man använder Chat-GPT för att lösa saker åt andra kunder. Vad det finns för möjligheter att göra det i säkrare miljöer än via free version utav Chat-GPT och liknande. Att man skapar en medvetenhet hos människan, nu när det blir en sån enorm demokratisering. Så tror jag.

Ludwig: Du anser alltså inte att vi behöver vara oroliga för den stundande AI-utvecklingen som pågår just nu då?

Josefin: Nej, det tycker jag inte att man behöver vara, för man ska vara glad för den. Men det är klart att vi har ju alltid, det finns risker med AI såklart och dem har ju funnits hela tiden och i många år och det blir på en annan skala men då är det inte framförallt det att det ska komma en terminator och liksom ta över världen, där är vi inte idag. Man tränar inte modellerna på det sättet men däremot finns det ju en risk för misinformation och fake news har blivit super sofistikerat just med både sättet att uttrycka sig och även skapa bilder som är väldigt trovärdiga i och med att de tränas på, i detta fallet internet och det är data som framtagit dem av människor med alla dem bias och liksom felaktigheter som kan finnas när människor uttrycker sig och göra saker så kan ju det förstärkas såklart, ännu mer än vad vi någonsin stått inför innan, att det så extremt många användare nu.

Ludwig: Tror du att vi kommer kunna leva i ett samhälle där AI fungerar autonomt eller tror du att det alltid kommer krävas en human in the loop?

Josefin: Jag tror att det alltid måste finnas en human in the loop, alltid. Även om man helautomatiserar saker och låter AI sköta saker så måste man hela tiden ha en övervakning på det och det är ju inte bara för att den kan springa iväg och göra knasiga saker utan det är för att man tränar på data som representerar den verkligheten som vi är i exakt nu sen att man ska ta ett viktigt beslut i samhället, för vi ser ju allt med att AI mäter kritiska beslut i samhället och då tränar man på data som är exakt nu för att det ska vara så relevant som möjligt men sen är ju verkligheten på något sätt relevant, jag menar det är ju bara gå tillbaka till pandemin

där allting ändrades väldigt mycket över bara en natt eller vad som händer i Ukraina, många modeller och data har blivit inaktuella på väldigt kort tid. Då måste man ha någon sorts möjlighet att gå in och se att det verkligen blir rätt, att modellen hänger med för dem ska kunna tränas om och ofta är det så att man har ett larm som säger att nu börjar det glida i korrekt att man tittar på performance osv och då måste man kunna träna om den så den verkligen är anpassad för den verklighet som vi har just nu. Det är super super viktigt. Och sen finns det ju system för det såklart som kan hjälpa till med det som människan i sin tur kan ha koll på, man kan såklart ta hjälp av AI för att monitorera AI men just AI governance som man kallar det för styrning och ha koll på den AI som vi har den börjar vi få allt mer och mer fokus på när vi ställs inför den accelerationen vi gör nu i samhället.

Ludwig: Om man tänker mer ett maskininlärt AI-system finns det något sätt att utveckla det i framtiden för att öka säkerheten och förtroendet för det?

Josefin: Det är ju framförallt transparens och förklarbarhet, historiskt sett, desto komplexare modeller desto mer black box har det varit, jag menar precis innan vi fick dem här språkmodellerna med miljarder parametrar så fanns det ju deep learning som hade miljontals parametrar iallafall och dem kan ju liksom inte bara slänga fram en lista med bara massa parametrar för någon för att förklara exempelvis om man fick ett lån eller inte eller vad det kan vara för något i slutändan utan där har man gjort jättemycket jobb senaste åren bara för att få fram metoder för att jobba med förklarbarhet hos modeller, det kan liksom vara en metod som lagts på efteråt eller metod som beräknas under tiden så att modellerna faktiskt blir förklarbara och nu är ju dessa stora språkmodellerna tveksamma för open AI som kanske egentligen borde vara mer öppna med tanke på deras syfte från början, dem har ju inte berättat hur det funkar och då är det svårt att säga hur transparent det är. Men övriga kända algoritmer, algoritmtyper för modellerna är liksom när man tränat en algoritm på data så finns det metoder för att förstå dem idag och det är ju såklart super viktigt. Sen är det också transparens i besluten såklart och transparens i hela kedjan, vad datan kommer ifrån och vad man gjort med datan som man tränat på, hur man manipulerat det eller kompletterat det och hur besluten tas baserat på resultatet som kommer från modellen och hur man väljer att hantera det utfallet som modellen spottar ut sig, så man har transparens i hela kedjan, då förstår man hur något fungerar och då är det lättare att lita på det. Sen har vi det hela med AI-act, på de kraven som är högrisk som EU:s förslag då, högrisk är den AI som påverkar individen hälsa, säkerhet, ekonomi och mänskliga rättigheter så dem grejerna är de som kan få en direkt påverkan på oss. Vet man att det ställs krav på dokumentation och rapportering och transparens osv på dem kraven som finns så vet man också att det är tryggt att använda det som har högrisk i samhället och det är ju hela förhoppningen med AI, att man ska kunna känna sig trygga med att använda den AI som finns i samhället, det är min förhoppning, att det faktiskt ska bidra till det. Sen är det ju jag menar AI-act är ju ett regelverk och ett ramverk men tanken är ju också att alla som vill placera produkter på vår marknad oavsett om de kommer från kina eller något annat land så måste dem ju då förhålla sig till våran AI-act för att få placera produkterna på vår marknad, vilket också skapar en trygghet för medborgarna och också såklart bidrar till vi får ett litet “schung” här i EU i och med att vi tvingar andra till

att också bygga in det i deras affärsmodeller, även om man kommer utanför EU. Och dem är kanske inte alltid jättebra på det utanför från tex Kina hållet.

Marcus: Jag tänkte på det att, tycker du att det är för stort fokus på själva utvecklingen gentemot utvecklingen av den här regleringen eller kontrollen över AI?

Josefin: Just nu tror jag att det skiftat väldigt mycket till att folk är rädda och att det är lite skrämsel och det har dem bidragit till med det här brevet kan man ju säga då även om intentionerna är goda eller så, det vet jag inte. Jag tycker det är lite svårt att veta vad intentionerna var med det, många har blivit väldigt skrämda av att systemen verkar så enormt sofistikerade på sättet som man uttrycker sig och sättet man kan resonera med dem. Så just nu är fokus väldigt mycket på liksom att det är farligt, inte så mycket på hur man ska hantera dem här riskerna och behovet av att hantera dem här riskerna utan folk är mer "vi måste stoppa det här innan det är försent" och det är ju beklagligt såklart.

Marcus: Jag vet inte hur mycket det uppmärksammats i dagen media men det har väl ändå framkommit där framförallt om Elon Musk säger någonting så står det ganska tydligt, du sa det där lite indirekt där kanske men tycker du att det skapar en oro som är eller inte är befogad lite åt den normala kanske inte jätte eller rättare sagt någon som inte använder AI så jättemycket för att dem bygger upp någon rädsla på det eller?

Josefin: Ja, det tycker jag. Det tycker jag är en onödig rädsla för vi är inte där idag så och det är inte säkert att det är syftet från dem heller men konsekvensen blir ju också en väldigt stor okunnighet i samhället och då blir det läskigt, så är det ju med allting som man inte riktigt förstår sig på.

Ludwig: Vi har varit inne en del på det nu, det här med förståelsen. Men när det kommer till riskerna, vilka typer av risker det finns med att fortsätta utveckligen av AI om det skulle vara så att vi tappar denna mänskliga förståelse och kontrollen för det? Så vi undrar väl lite, vilka risker som egentligen finns?

Josefin: Dem riskerna som jag ser är de risker, det finns risker med alla AI-system, att de hamnar i fel händer och att det blir någon acceleration av missinformation eller acceleration utav bias eller diskriminering osv. Dem bitarna som vi haft hela tiden egentligen och som kräver ett stabilt governance system och som kräver transparens osv. Alla dem bitarna. Det är ju, för man har ju hela tiden kontroll över datan som tränas på, jag menar GPT-3 exempel den har internet fram till 2021 eller vad det är och nu så lägger dem på mer, det är ju inte det att dem springer iväg och gör andra saker men däremot API:er som folk kan komma åt och bygga in i egna system och bygga saker som det inte va tänkt att användas för.

Josefin: För att använda det på ett felaktigt sätt och det är snarare där vi behöver fokusera på att se till att komma åt sånt, och där blir också transparensen mindre när folk kanske använder en api för att ta fram saker utan att riktigt då kunna ha insikt vad som sker i själva grundmodellen som man har exponerat, som ger möjlighet att bygga in i sina egna

applikationer. Den som byggt applikationen, som liksom använder api är inte säkert att den personen har transparens i, insikt i hur själva modellen egentligen fungerar. Dem riskerna blir ju värre likaväl som om det börjar komma ut någonting knasigt som får spin, så man börjar ta, ah som man börjar sprida helt enkelt felaktiga rykten eller felaktiga nyheter och sånna saker är värre.

Ludwig: Sen var vi lite inne på det här med att liksom hur man ska kunna lite på det resultatet AI genererar utan att förstå hur datan är insamlad, eller vilka AI regler som egentligen är tillämpade under utvecklingen. Och då är lite såhär, hur stark tillförlitlighet känner du kring det resultatet AI kan generera idag? utan att veta om det är liksom hur den här processen egentligen har gått till.

Josefin:Om man säger den AIn som jag själv arbetar med, ifall jag skulle vara med och utveckla ett system , då känner jag tillförlitlig till det och den AI vi har i samhället idag. Så jag tycker som tidigare sagt, där har vi inte insyn i exakt hur det här skapandet funkar men vi vet att det baseras på reinforcement learning och deep learning. Vi vet liksom att det är inget märkligt som verkar i bakgrunden, men däremot vet vi också att det fortfarande är på forskningsnivå så man, det som kommer ur den kan man se är rena felaktigheter, att den liksom tar var den har om man säger så. Jag tror det snarare handlar om att vi måste kommunicera vikten av att kritiskt värdera det som man får ut så man inte bara köper det med hull och hår utan mer ser det som du sa innan “human in the loop”, att man ser det som ett verktyg, ett komplement. skulle jag skriva ett examensarbete tillexempel så hade jag självklart använt Chat-GPT för att få nya idéer, nya vinklar och sätt att formulera mig och uttrycka mig på ett bra sätt och hur man skriver en avhandling, asså jag hade bett den.

TAPPAR KONTAKTEN

Josefin: Jag sa nog mer att jag skulle använda det och jag skulle liksom. Och Jag använder det, verkligen, i jobbet, det gör jag. Men man måste kolla referenserna, man måste kolla upp så att det stämmer, men man kan se det som en starter, som om jag tillexempel skulle ha skrivit min avhandling nu istället för 15 år sen så hade jag kanske hittat många fler roliga referenser och jag hade kunnat uttrycka det här som var ganska, man blir liksom ganska nördig när man forskar om någonting och vokabuläret blir svårt för någon som är utomstående att förstå, då kan man få hjälp att skriva det på ett bra sätt och det är ju ingen nackdel med. Då för man ju fram sitt budskap på ett helt annat sätt. Man kanske får lite andra infallsvinklar som man kanske inte har tänkt på som man kan grotta vidare i. Men sen när man examineras och när man står där mot sin opponent, så hade det hade det varit otroligt mycket viktigare att lägga vikten på: vad tror du, varför valde du den här metoden och inte den och vad tror du det här kommer ha för påverkan på det är och det här området om 5 eller 10 år och vad har det här som inte det här eller det här arbetet har. Liksom sånt har resonerande som inte en AI kan hjälpa till med, som är mycket mer din egna kunskap. Och det man, det man kommer med när man gör sitt egna arbete det är någonting nytt som inte finns att tillgå i dessa systemen redan.

Marcus: Många har varit inne på det här med att AI kommer ersätta många arbeten. Är det någonting människan behöver vara oroliga över tror du?

Josefin: Jag tror att det kommer säkert ersätta en del arbeten, i alla fall hjälpa vissa arbeten att bli mycket mer effektiva. Jag tror inte man behöver vara oroliga över det för historiskt sett har man sett när, när det kommit nya möjligheter då har också arbetsuppgifterna utvecklats så det brukar ju snarare vara en revolutionär fördel för att samhället ska gå framåt. Det är många som säger att man ska hitta meningsfulla mänskliga uppgifter och så kan det ju också va men jag tycker vi ska se en utveckling att vi avancerar och får liksom en helt annan output att göra saker snabbare och jag tror att det bidrar till utvecklingen jag tror inte det kommer bidra till att folk blir arbetslösa. Tvärtom att man slipper att göra de där knasiga sakerna man kan låta en maskin göra åt en, det är bara bra.

Marcus: Vi läste en artikel som du har skrivit, där det står att ni jobbar främst med att skapa medvetenhet och vägledningen till organisationer i hur de på bästa sätt kan implementera AI ansvarsfullt. Vad är den främsta anledningen ni har stött på, vad brukar vara en organisations syn på AI, varför de inte implementerar det eller kan vara lite motstridiga mot det?

Josefin: Jag tror ganska länge så har det handlat mycket om vad AI är. Man fastnar lite i definitionerna *vad det är för någonting* och *vi gör ju statistiska modeller, det är ju inte AI*. Man tänker att det rör inte oss för vi håller inte på med sånt. AI det är såhär någonting som bara är adaptivt och superkomplext och det är läskigt. Ganska länge och fortfarande till viss del så var nog det anledningen till att man tänker det inte är för oss för vi gör andra saker. Nu tror jag de allra flesta organisationer som jag pratar om vill ju använda AI, inser att de måste använda AI, man kan inte komma och säga att vi jobbar inte datadrivet för då är det lite kört konkurrensmässigt. Men då ligger snarare utmaningen i kanske systemen, om man har saker på plats att verkligen kunna använda, få saker att fylla en funktion, att få saker att bygga modeller, man kanske tittar på vilka kunder ska nappa på ett erbjudande eller vilka kunder som ska säga upp sitt abonnemang osv. isolerade modeller som någon sitter och kör gång efter gång i batch. Men att verkligen få det produktionssatt så man kan börja ta beslut på en större skala eller börja göra det i ett affärssammanhang där man sen genererar använder det som kommer ut, inte bara spotta ut där man sen funderar på vad man ska göra utan koppla det till någon action som faktiskt leder till leder till en affärsvinst eller affärsförändring eller att någonting händer, att komma dit är nog snarare en mognadsfråga som många vill ha hjälp med att hitta dit och förstå vad man behöver för att komma dit. Det kan vara att det sitter ett par datascientists som bygger saker i open-datasource, vilket inte är något fel med men dem modellerna som dem är har byggt i (fighten?) dem behöver kopplas med ett beslut och dem behövs kopplas med data, effektiv dataprepp och dem behöver kopplas, tränas på nya data kontinuerligt så man får hela grejen att snurra där. där ligger utmaningen att det inte bara blir isolerat experiment utan att det faktiskt blir något som är produktionssatt. Då är ju mitt team, jag vet inte vilken artikel ni läste, men där vi framförallt vill ha in i den diskussionen är ju då att man sätter upp det och faktiskt får det att snurra då måste man också se till att ha möjlighet att förklara besluten att man tittar på fördelningarna på datan innan dem går in i,

när man tittar att man inte har känslig data och man följer upp beslut så att de är rättvisa och hela, att man har koll på de grejerna under hela kedjan.

Marcus: Vad tror du den, vi har redan varit inne och snuddat på det men den värsta tänkbara konsekvensen av dagens utveckling? och då tänker jag inte på, eller lite på det här GPT uppproret

Josefin: Jag tror, värsta tänkbara, jag tror att vi är i ett läge där det kommer bli farligt för mänskligheten på det sättet som folk är oroliga för. Det handlar nog snarare om att det kommer personlig information som felanvänds så att det blir privacy breaches till exempel, ni så den här, det är en liten breach där en kort stund där folks personliga data exponeras, sådana grejer får ju helst inte hända och det är sådana grejer som kan drabba individen helt enkelt.

Josefin: Och som sagt då, när det blir, när det hamnar i fel händer och i stor skala skapar falsk infomration och det kanske inte är jätte stor risk just i Sverige men det finns ju andra länder där man kanske inte har samma öppenhet på internet tex, där man inte har samma möjlighet till omvärldsbesvakning och kan då bli utdatead. Tex jag menar hade vi vart i ett land där vi inte har tillgång till omvärlden och internet liksom som Ryssland tex då hade man kanske tänkt att amen tex den där påven med dunjackan. Man kanske inte hade tänkt att det var påven, de kanske inte hade varit någon fara men han kanske inte alls vill bli igenkänd som rappare eller vad va det?

Wictor: Aa precis

Josefin: Det är lätt att sprida, det är rätt lätt att luras och det kan drabba individen och det kan göra det på en stor skala om man inte får till att verkligen få säkerheten på plats i samband med det här. Då menar jag liksom inte säkerheten med att springa iväg och göra saker på på, farliga saker utan att man har förklarbarhet, att man har transparens, att man har ett etiskt filter liksom på det som kommer ut.

Marcus: Juste. Det här öppna brevet, tror du det är någonting som kan ha negativa effekter om man skulle införa en paus på 6 månader, minst 6 månader när det kommer till själva utvecklingen och oron i samhället och sådär?

Josefin: Jag tror liksom, skulle man pausa 6 månader skulle det inte göra någon skillnad i det här läget och för det första så kommer dom ju aldrig kunna säga till vissa länder att sluta, liksom det är ingen som skulle bry sig eller göra det och sen jag menar, kommer väll dom ifatt senare då men den skada som det gör när man säger det, det är snarare signalen det ger att dethär är läskigt liksom, det är nog snarare en skada. Det är inte bra att man skapar en oro för det här är någonting som också kan, inte just jaa… repertati också men att AI har ju möjligheter att hjälpa männskligheten på många bra sätt och det är ju en bra utveckling och framåt. Man hade kanske inte kunnat tro för ett år sen att vi skulle vart där vi är idag, det hade jag ju aldrig trott om nån hade frågat mig. Det har ju gått väldigt väldigt mycket fortare än vi hade tänkt men på ett positivt sätt.

Marcus: Ska jag se om vi kunde hitta på något mer vi kan fråga här.

Wictor: Jaha, var det sista frågan?

Marcus: Nja, du har ju svarat på ganska mycket här redan så att…

Ludwig: Men jag tänker, du har ju pratat väldigt mycket om möjligheterna och fördelarna med AI.

Josefin: Mm

Ludwig: : Vad tänker du, alltså hur långt tror du vi kommer vara om 5,10 år med den här AI utvecklingen som, om den fortsätter gå lika fort som den gör idag? Alltså hur det hjälper oss egentligen?

Josefin: Ja, det är svårt att veta. Jag tror att det kommer gå ganska snabbt framåt nu och såklart på ett positivt sätt också men risken är ju som sagt… Jag pratar ju väldigt mycket om fördelarna och vad det kan ge men jag pratar även väldigt mycket om vikten av att vara ansvarsfull och att det ska vara tillförlitligt osv och lägger vi ingen fokus på det, liksom inte komma någonstans på den biten då blir det ju, hoppas jag vi inte kommer jätte snabbt framåt på utveclkingen häller för den behöver ju vara, det behöver ju vara en balans mellan dom 2. Jag tror att man allt mer kommer se, som en individualisering utav vad man kan göra med AI, mer hjälp på individ nivå, liksom säg inom sjukvården att du kan hitta mycket mer saker preventivt innan det blir farligt. Tex Sjukvården eller i trafiken eller i konjunkturen eller i lagerhållning hos en butik, liksom allt man på nåt sätt mycket tidigare kan se vart de bär och vad man kan göra snabbare. Där tror jag vi kommer se en stor nytta, särkilda förhoppningar inom hälsa & sjukvård men även alla dom här komposerande som vi ser i alla dom här språk modellerna och den softisikeringen som man ser, det tror jag också kommer kunna liksom bidra till väldigt mycket till äldre som behöver konversera för att… som är ensamma. Jag tror att man kommer kunna se att AI kommer kunna ha en stor påverkan på tex vad gäller klimatforskning och om man verkligen bara använder det på ett bra sätt finns det så mycket vi kan göra idag som vi inte gör redan med AI som också skulle kunna ha otroligt mycket bra fördelar för männskligheten och för stora problem. Men det är som sagt, det krävs att man inte är rädd för AI utan att det krävs snarare en ökad kompetens och en, en… ah att man litar på det. Det är ju det ena, att man jobbar med riskerna och att man jobbar med att det faktiskt finns… kommer en reglering så att folk kan känna sig trygga, då kan man också komma framåt för att finns det "Trust" finns det också möjlighet till innovation.

Ludwig: Men har du någon egen liksom erfarenhet kring när det har gått åt andra hållet? Som du skulle vilja dela med dig av… Om det är något du kommer på just nu liksom?

Josefin: Nä inget projekt jag har varit med i och inte egentligen någonting som SAS häller varit involerade i veterligen i något projekt så. DEt har ju finns tillfällen där liksom SAS

programvaran har varit med där det blivit fel liksom så då. Men själv har jag inga sådanna erfarenheter snarare bara att när man jobbat med modeller har man identifierat att det här var inte rätt för att här påverkades vilket… den som skulle bli vald till denhär studien påverkades utav postnr och det får det inte, då får vi tabort det så får vi kika på vad mer som är relaterat till postnr eller det här ska inte en tjänst vara med och betämma tex eller att ålder blev viktigt här, det kanske inte är okej och även tvärtom att man har fått diskutera exempel om ursprung inte ska ha någon betydelse, i visa fall är det jätte jätte tabu men i vissa fall är det jätte viktigt för att man ska få rätt vård tex för man har olika förutsättnignar och olika genupsättningar sådär. Att det mer har varit resonomang kring så att det ska bli rätt och SAS nu jobbar jag ju på SAS och vi har ju en AI plattform där man har funktionalitet från data till beslut för att kika på hur det blir, så att det blir tillförlitligt över hela kedjan så att det är snarare att man har hittat något fel på vägen innan man sätter det i produktion och att man också har möjlighet att få larm sen när saker blir knasigt så man kan trycka på paus eller träna om, och det behöver inte vara något drastiskt utan det kan vara att nu blir det fel hela tiden, hon hade inte alls tänkt säga upp sitt abbonemang. Man måste liksom sluta ringa till dom som redan är nöjda. Men det finns även väldigt många exempel på när det blir jätte fel och det är ju det klassiska exemplet när föreslår bättre jobb till män eller när man dömmer svarta hårdare än vita eller ger kredit till, högre kredit till män än till kvinnor fast dom får samma kredit score fast det allra allvarligaste i nutid är tycker jag nog var det här, var det i Holland var det va, Nederländerna där man hade ett system där man där man tittade på försäkringsbedrägeri. Det var väll bidrag till hemmen som hade ekonomiska problem och den modellen dömde ju fel och krävde återbetalning, det har ni säkert läst om, för väldigt många människor och det fick pågå under många många år som då ledde till att väldigt många människor fick, alltså att de fick betala tillbaka väldigt mycket pengar fastän att, det visade sig sen att den här modellen var diskriminerande där man felaktigt straffade dom som kom från vissa etniska minoriteter och som då hade lägre inkomst och folk gick i konkurs liksom ekonomiskt och det var barn som hamande på fosterhem och det var skiljsmässor och det var självmord tom bland det här som kom upp och hela regeringen avgick ju. Var de 2019 tror jag till 2021 det uppdagades 2019 iaf, det är ju en ganska ny grej liksom.

Marcus: Är det en konsekvens på att det inte är tillräckligt genomskinligt eller transparant och…

Josefin: Helt klart! Hade vi haft tex EUs regelverk så hade detdär aldrig kunnat få hända så det är en typsik sådär grej man förhoppningsvis kommer att kunna motverka om vi har krav på att man måste kunna redogöra hur modellerna funkar och rapportera och registrera och eh ja osv. Så det är positivt om vi kan få det på plats så snart som möjligt.

Wictor: Vi hade en intervju förra veckan kan vi ju nämna och då var det väldigt negativt till AI utvecklingen och han vi intervjuade då pratade om människans undergång lite mer

Josefin: Haha, han har snackat med Tegmark och…

Ludwig: Aa men det är väldigt kul att vi får se olika synvinklarna vi får nu.

Wictor: Aa det är verkligen uppskattat att vi får olika synvinklar.

Josefin: Ja vi hoppas att vi kan vända på det nudå, lite gladare

Wictor: Tänkte hade vi några frågor mer eller var vi nöjda?

Ludwig: Det var nog allt vi faktiskt behöver.

Wictor: Aa

Josefin: Ja va bra, ni får höra av er om det är något som var oklart eller om det är något ni saknar.

Wictor: Ja absolut!

Ludwig: Ja tack så mycket för att du ställde upp.

Josefin: Kul lycka till nu, när ska ni vara färdiga?

Ludwig: 15de är la första utkastet men andra juni är slut inlämningen

Josefin: Är detta en magister uppsats eller vad är det?

Ludwig: Nä kandidatuppsats

Josefin: Mm, spännande, kul ämne

Wictor: Aaa det känns som det passar bra nu

Josefin: Ja, aktuellt. Får ni använda chat GPT och skriva ihop

WML: Hahahaha

Wictor: Nä men tack så mycket så återkommer vi om vi har några mer frågor

Josefin: Yepp tack så mycket

Wictor: Yes hej hej

Josefin: Hejdå

9.3 Bilaga 3: Transkribering från intervjun med Mathias Lanner

Mathias: Jag heter Mathias Lanner och jag jobbar inom det här området på SAS Institut. Vi kallar oss följt för Advisors eller Pre-Sales men experter, det kan man väl också säga. Jag gick för länge sen. Jag gick ut 1998 på något som heter Statistikprogrammet i Linköping. och har sen dess jobbat med analys. Jag har varit anställd på banker och konkurrenter, men jag har varit på SAS nu i snart 19 år. Jag har jobbat inom olika branscher där man använder sig av data för att kunna fatta kloka beslut. Då är ett steg på vägen kring modellering och att kunna skapa prediktioner som man sen kan agera på. Så det är jag. Ni har ju träffat min kollega Josefin, så vi sitter i samma team. Vi jobbar nordiskt här på SAS. Det är ett gammalt bolag som har hållit på länge inom område dataanalys men nu säger man mer att det är allt jävla skit i AI. Den definitionen är väl lika många som det finns människor som har åsikter om vad AI är. Men det är det all in come. Men kör igång så ska jag göra mitt bästa.

Ludwig: Jajamän. Ja, vi kan börja med första frågan då egentligen. Den är rätt bred men den pratar om vad dina tankar är kring osäkerheten och kring den stundande AI utvecklingen som vi har i dagens samhälle? Med tanke på att allting går väldigt fort fram just nu.

Mathias: Jag tänkte att något sånt här skulle komma. Om jag börjar med att svara som pappa, jag har två döttrar så är det jädrigt spännande, helt plötsligt hur den här utvecklingen, bara från egentligen november, december slår sig så brett mot undervisningen. Det finns fördelar och nackdelar att kanske snabbt samla in information. När jag gick i skolan hade man Wikipedia... Inte Wikipedia, det hette encyklopedin. Man hade stora feta böcker. Det kanske inte ni har haft. Man slog upp och sökte. Nu finns det att man kan köra på webben. Det är ytterligare en nivå att undervisningen kan påverkas ganska mycket av den här tekniken. Det är man inte redo för. Man märker att vissa lärosäten förbjuder det, det här får man inte använda när man gör inlämningsuppgifter då sitter man i ett klassrum och inte jobbar flera veckor i kapitim i det hela. Det jag tänker då är att det här måste man ganska snabbt komma underfunn med. För tekniken kommer inte att sluta fungera utan den kommer utvecklas. Så det här är det första jag tänker på. Hur ska det tas in inom utbildningsväsendet? För det är en sådan möjlighet och det kanske har varit hot också. För att saker man ibland får fram. Jag har inte jobbat speciellt mycket med ChatGPT, men man hör ju mycket. Det kan ju vara felaktigheter. Det är information som inte är eller på något sätt, den är ju något föråldrad, kan det vara. Men i vissa fall är det högst relevant. Om man tittar på sevärdheter i New York är det inte att de ändras så mycket från dag till dag. Så vissa saker är relevanta, men i vissa saker kan det bli felaktigheter. Men det är också som jag tänker att det här kan på nåt sätt... alla blir så likriktade man får ta tag på samma information. Det är bara att formulera en fråga. Sen får man fram ett svar som man på nåt sätt tycker, det främjar inte egen tankekraft. Det kan vara innovationsdödande just för de yngre människorna där det är allt så digitalt. Man skriver och det kommer ett svar som bara är perfekt. Jag tror man kan bli lite dum och lat. Det är nog det jag vill lyfta fram och ta död kanske ibland på kreativitet. Kreativitet ska vara att kunna använda detta ihop med annat. Det är en sak som jag tänker, först och främst, eftersom jag har barn som en går på gymnasiet och en på skolan, där det här har blivit. De har ju själva

använt detta. Det har varit ett sätt att, om man ska göra en muntlig framställning och prata om olika resmål i New York. Då är det skitsmidigt att fråga ChatGPT vilka de mest kända resmålen är än att själv ha varit där och börjat grotta fram detta. Men om de används i uppsatser i bedömningen är det bra att kunna lägga ner kraft på det retoriska hur jag istället ska formulera och agera. I ett sånt läge kan det vara bra, men om man ska göra ett examensarbete och få jävla mycket hjälp, så måste man tänka hur man använder det och vad man ska bli bedömd på. Jag är inte helt klar i vad jag tänker, men man behöver hitta nåt förhållningssätt. Jag har inte svar på det, men det här behöver man fundera på. Ni har ju säkert snackat om det jättemycket i Borås. Så det är en sak som jag tänker på. Sen är det ju om man tänker sig inom kanske företag så kan det vara lätt att man trycker iväg mycket information som man kanske inte borde trycka iväg. Om man tänker sig vi på SAS, det här verkar ju smidigt och sen så matar jag in lite kod. I den koden kanske det fanns nåt login till en maskin som kanske någon annan skulle kunna komma åt. Så att det finns en fara. Vem äger detta? Jag antar att det är ChatGTP, eller motsvarande. Det är de som äger den informationen som man stoppar in i det hela. Det får de göra vad de vill med, och det är också en risk. Sen är det ju om man tänker sig att vissa jobb kommer ju att kunna ganska snabbt försvinna, och det är ju en fara. Man vet ju det här med digitala kameror. När jag var ung tog man film och lämna in på framkallning. Då var Kodiak störst på det hela. De är ju inte i den branschen. Jag läste nånstans att det var ett företag som jobbade digitalt med läromedel och att kunna hjälpa ungdomar på något sätt och deras börskurs bara försvann. Så det kommer att ske i det här fallet att företag bara går under för att tekniken är så stark och tar över det de har jobbat med. Det är en stor risk. Men det finns även risker med allt som handlar om att producera text och sammanfatta information och allt möjligt. Hela den biten kan jag se i nuläget. Sen är jag kanske inte så rädd för att de ska ta över, att maskinerna blir klokare än vi. Jag tror inte jag har den riktiga målbilden. Jag hoppas att vi människor ska kunna vara kloka där och inte låta det skena iväg. Men det är väl bara i min positiva människosyn. Man brukar prata om singulariteten när det här tar över oss. Då är det ju bara att rycka ut sladden då. Jag brukar säga att det behöver ändå ström för att det ska fungera. Det är ett korkat begrepp. Men just nu är jag trygg i vad jag sysslar med att jag inte ser riktigt att jag målar fan på väggen. Eller så kanske man skulle göra det, för jag kan ha för lite kunskap. Det är ungefär så jag känner och tycker just nu. Men det kan komma fler frågor. Men det jag tänker är att när det här kom så har begreppet AI blivit på varemans bord. Så det som har hänt från november till nu, det har ju hänt mer än under mina 25 år i arbetslivet. Om man tänker gemene man tänker kring vad man kan göra med dataanalys. Det är en teknikutveckling som är helt sjuk just inom det här området som alla på något sätt kan ta till sig för vem som helst kan skriva en fråga och få fram något som är skitbra, detta är ren magi.

Ludwig: Men anser du att vi behöver vara oroliga över den här teknologiutvecklingen som vi pratar om?

Mathias: Ja, jag tycker att vi behöver vara, absolut försiktiga. Oroliga det kan man säga att det är en synonym till försiktiga men vi behöver vara försiktiga. Vi måste få till kloka regelverk som är världsomspännande. Vi har regelverk som kanske kommer inom EU. Jänkarna kan vara liknande. Men man behöver hitta nåt sätt att det inte skenar att man

använder klokskap för att kunna hitta nån form av kontroll som inte är innovationsdödande på något sätt. Jag tror inte heller att bara nu ska vi sluta med detta. Hitta en enighet hur man gör det här på ett klokt sätt. Det kan inte jag svara på. Jag tror inte på att bara stopp och belägg och nu är det slut. Förbjuda kommer inte att göra något bra, för det kommer inte att kunna gå att förbjuda nåt sånt här. Men jag är lite orolig, absolut, framför allt om man inte får till ett sånt här kontrollfunktion eller motsvarande dodo absolut. Sen är jag orolig för att det kan ta död på våra barn. Ni kanske inte har ungar, men deras kreativitet och kunskap. Alla gör ungefär samma sak. Vi blir mer likriktade och har samma information och blir dumma ochh lata det är jag livrädd för. Att man blir för bekväm och bara skriver svar och får fram svaren och är nöjd med det. Inte hur svaret togs fram, utan man bara köper det som kommer fram. Det är ungefär som att jag brukar titta på mina ungdomar när dem tittar på youtube. Jag tycker att de youtubers är sminkade och packar upp varor och är glada liksom. Det tycker mina smarta tjejer. De blir föredömen. Det tycker jag att nu ser vi att alla vänder samtidigt. Det är jag rädd för att det blir nåt liknande. Man blir mainstream och tappar sin kreativitet och köper det som kommer.

Ludwig: Lite att man tappar den mänskliga kunskapen då.

Mathias: Ja. Den kunskapen som man alltid kommer att få är något som redan finns. Men det främjar inte innovation och kunskap eller tänka i nya banor. Barnen blir bekväma. Det är snabb respons, du får nåt svar. Det tycker jag är läbbigt.

Marcus: Anser du att utvecklingen är utom kontroll? Om man saknar nån typ av begränsning eller att man ska lära AI på ett speciellt sätt?

Mathias: Det är svårt att svara på den frågan. Om det saknas en kontroll hur man tar fram det här just nu, är det som var frågan?

Marcus: Ja, men precis.

Mathias: Det gör vi absolut inte nu. Här har det kommit nya versioner senaste tiden som talar om vilken information de har tränat på. Men jag tänker att man behöver lägga på ett lager som kan verifiera det som kommer ut, de dumheter. Det finns massor av exempel när folk har involverat sig i såna här chatt-funktioner och det blir tokigheter. Så att lägga nån form av funktion över svaret att kunna göra en validering, att det här är korrekt och följer etiska normer och riktningar att få fram ett sånt lager över det som spottar ur. Det behövs en samsyn på vad det ska kunna vara. Där kanske vi har någonting som kommer från EU. Men här behöver man tänka att det blir i ett större sammanhang. Att man likriktar även för säg kineser eller ryssar men att man tänker stort och brett och tänker på människans bästa om det ska låta bra.

Ludwig: Men tror du att vi i framtiden kan leva i ett samhälle där AI fungerar helt autonomt? Eller bör det ändå finnas människor som övervakar och kontrollerar de här AI-systemen?

Mathias: Ja ja, jag är 53-bast och har det svårt att se det där framför mig. Vad ska vi göra? Ska vi sitta här och kolla Netflix? Jag skulle gärna vilja se att det inte är helt autonomt. Jag vill att det ska finnas en interaktion. Fortfarande kan människan tänka andra saker som man inte systemen kan göra. Att man pratar om en augmented… Att man ska dra nytta av båda aspekterna. Sen om det finns AI som kan göra allting. Det är vi inte riktigt där som är multidimensionella och kan göra mer saker än vad det kan idag. Jag tror på människorna, att det alltid måste finnas en symbios tror jag, absolut. Låter det bra?

Ludwig: Ja, det gör det. Hur tror du att man hade kunnat utveckla de här AI-systemen för att öka den här säkerheten och förtroendet för systemen? Liksom att det ska bli mer acceptabelt för människan i samhället?

Mathias: Jag tror att för att det ska bli mer acceptabelt så måste man dels kunna poängtera att det kan bli fel. Det som kommer fram är inte alltid rätt. Våra ägare på SAS frågade systemet vad mina döttrar gör. Den ena var utvecklingschef på SAS, men det var hon inte och den andra var en framgångsrik kock så hade jag ställt en fråga och inte vetat det hade jag tagit det för givet. Man måste nog poängtera att det här inte är sanningen. Det här är nån form av... Att förklara för människor att det här är en prediktion om... Det här är text som genereras utifrån nåt jag skrivit in. Man har tagit fram att det här borde vara ett rimligt svar på frågan. Att verkligen tala om att det inte är en sanning, utan en beräkning. Sen kanske man nu kommer på saker att man ska kunna ha någon form av relevansscore att säga. Det här har vi sett liknande och att folk kan föra in en feedback i sina svar. Det här var bra, det här blev dåligt. Jag vet att det kanske finns i vissa av de här men jag har inte använt det här så mycket. Men att det finns en feedbackscore där man hela tiden kan föra in att det här är schysst, det här var bra och sen kan man kunna bli belönad om man använder det här mycket. Att du blir rewarded om du får in bra saker och det här har varit schyssta grejer. Att det kan belöna. Ju mer relevanta svar som föds in så blir systemet bättre. På något sätt att vi alla som använder det kan hjälpa till och tala om att det här var inte bra. Då kanske man inte skenar iväg i sin träningutveckling. Nu glömde jag bort frågan om vad som skulle krävas för att utveckla den på ett schysst sätt.

Ludwig: Ja men liksom hur ska man utveckla AI-systemen för att öka säkerheten och förtroendet?

Mathias: Ja men då tror jag att dels att gemene man måste utbildas. Förut har vi som har jobbat med det här länge företag som använder våra system som vi säljer på SAS. Det är inte att jag har gått och sålt till min mormor- eller köpt ett neuralt nätverk. Man måste utföra… få upp kunskapsnivån. Då kan jag tänka mig att det här måste ske tidigt i skolan- när det här blir så tillgängligt. Att tala om det på ett bra sätt, göra sköna, instruerande filmer- som barn kan tycka om, då kanske youtubers kan ha en jävla viktig roll att det är risker med det här. Det andra sättet är att få till nån övergripande tillsynsmyndighet eller motsvarande som på nåt sätt kan övervaka detta. Man kanske kan ha regelbundna avstämningar. Hur långt har man hunnit? Vad kan systemen kan göra. Man måste ha kontroll så att ingen kanske smyger i buskarna och tar fram nåt som inte alls är bra. Det kanske kan vara att man ska leverera ett

system som går in, nu kom jag på något bra. Det är som läkemedelsindustrin. Om man ska få ett läkemedel godkänt följer man en jätteprocess med kliniska prövningar i olika steg och sen måste den godkännas. Sen måste man använda mjukvara från SAS, för den räknar på många decimaler. Men kanske en sån process för att få så godkända att gå igenom en sån kontrollfunktion ...det kanske skulle vara någonting, att man tittar på saker som kan påverka en som läkemedel. Det kan påverka oss om det är dåligt. Där har man fått till en regulatorisk process. Kanske nåt liknande här som är övergripande. Mer än bara AI-act där man tittar på vissa saker utan tar fram hur man har tränat det och hur man har gjort. Det kanske skulle vara nån bra sak.

Ludwig: Det låter bra.

Marcus: Vilken typ av risker ser du med fortsatt utveckling när man inte har nån mänsklig förståelse eller kontroll?

Mathias: Riskerna är dels, som jag har sagt, att jag tycker att det på nåt sätt tar bort innovation och såna saker. Sen finns det en risk att man tycker att om man skulle förbjuda det- så tar det bort innovationen. Det är på båda sidorna. Men sen är risken att när människor blir för involverade- och tror att detta är sanningen och lever sitt liv ihop med en chatbot. Att det kan få verkligt hemska konsekvenser. Att det blir självmord och såna tokigheter. Sen är det kanske ännu mer risker om det här skulle skena iväg på nåt sätt och kunna kapa en jävla kärnvapenterminal. Eller vad det skulle löpa mot en sån här system och kunna bli så smart att det kan manipulera… där är vi ju inte riktigt än men det är väl såna risker man ser, att det här kan vara människans undergång. Om man tittar på dystopiska scenarier. Jag ser väl inte det. Men om kloka människor har sagt att det finns en risk så finns det absolut det här att systemen vill överlägsna oss. Det finns en bok som har skrivits av Dan Brown. Han har läst den? Det handlar om att det finns en snubbe som har skapat… Boken kom för sex, sju år sen. Han är långt före Chat-GPT men han vill nå ut med sin kunskap. Han har skrivit en bok. Han är inom konstvärlden. Han har en A.I. han kan snacka med hela tiden. I princip så händer det att han dör, den här snubben. Han är skummis, men han dör för det bästa sättet att få publicitet är att han blir skjuten eller dödad- på sin vernissage av den här boken. Men då är det hans AI som har iscensatt detta för att nå maximal publicitet. Sök på Dan Brown. Den kom för ett antal år sen, men den är super relevant nu. Han låg ett antal år före, men då har AI blivit så smart att han inser att det här är bästa sättet att få ett stort genomslag. Det är det där som är lite läbbigt om det kan komma på den nivån. Men har vi de här regulatoriska processerna så kanske man kan... Det går inte så pass långt.

Marcus: Hur ska man kunna lita på de här resultaten? Det var det du var inne på det lite…

Mathias: Aa men där tror jag att lägga någon typ av kontrollfunktion som kan checka mot relevant kunskapsbas

Marcus: För vi har sett… Vi har haft en annan intervju där dom tog upp tex hur Chat-GPT kan lägga till referenser eller källanvisningar…

Mathias: Ja, men det har jag sett. Dels källhänvisar så att det är lätt att komma på. Sen är det ju det här att kunna få nån form av... Kanske har andra frågat detta. Det här har styrkts så att det har varit bra svar. Kanske att det skulle kunna vara nånting. Men sen också att... Ja, det var nåt mer som jag tänkte på, men nu kom jag av mig. Utifrån vad man frågar om vad systemet används till så borde man ha nån form av risk inom området. Om det är att jag är personen med systemet så är det kanske en annan risk än när jag frågar om trevliga resmål i Rom. På nåt sätt har där en form av riskindikator utifrån vad jag ber systemet om. Det är en känslig sak. Om jag skriver in en känslig fråga som handlar mer om mig själv så kanske jag kan bli innan svaret kommer att det här kanske är ett svar som inte alls är relevant. Det är baserat på min lösa antaganden att ta detta inte för givet, utan det här är nånting som är simulerad text på nåt sätt. Att utifrån vad frågan ställs ge en form av tydlighet vad det här svaret kan innebära.

Ludwig: Yes

Marcus: Det här öppna brevet och du läste det med Future of Life som har kommit ut med pausen på 6 månader. Tror du att det är nånting som behövs eller kommer det bara hämma utvecklingen? Eller är det någon propaganda-grej?

Mathias: Jag tror att i det här fallet är svårt att stoppa det. Jag tror att det är svårt att komma med ett förbud. Hur ska det egentligen kunna se att folk som sysslar med detta följer detta? Jag tror att det är väldigt svårt. Sen sex månader, tittar man på vad som hänt på sex månader från nu till november så har det hänt skitmycket. Sex månader framåt, skulle man ta en paus, så är väl alla i startgroparna. Då kan vi få en dubbel utveckling nästa. Jag tror inte att det här har nån direkt betydelse. Man brukar säga att om man ska göra det i västvärlden så kommer väl andra i östvärlden att göra inte en paus. Så jag tror nog inte att det har... Men det är kloka människor som har sagt det. Jag anser att jag inte är så pass klok och att jag inte kan fatta ett sådant beslut. Men jag tror att det är svårt att stoppa det om man skulle ta en paus. Det kommer ändå att finnas utveckling, men man måste ta med respekt de här personerna som säger det. Det är inga dumskafft, utan det är kloka personer, så det finns någon relevans i det hela. Men jag tror att ett förbud är farligt. Det har man sett i andra. Att man förbjuder nånting... Vi sniffar ju ändå även om det är förbjudet. Jag tror att det är svårt med ett förbud, även om det kanske skulle vara bra. Men jag är lite svävande på frågan. Som ni hör, det finns både... Det kanske är bra och dåligt. Men förbud tror jag inte direkt på i det här. Vi kanske kommer på något supersmart. Eller den här personen som inte fick och sen kanske hon eller han blir påkörda så att vi missar det.

Ludwig: Du tänker ändå att det kan bli något bra utav det hela?

Mathias: Jag tror att det har varit, om man tänker mänskligheten, att det har varit... Nu är det ett ganska tufft läge i världen, men jag har fått känna hopp. Jag hoppas att alla människor kan hitta nån form av... Nu är jag lite flummig, men att mer gå på empatiska och såna saker. Att få fram hela sin vilja, ha totalt härarvälje. Jag tror att man behöver åka på några pumpar och sen

kommer det att komma... Människan behöver kanske lära sig också. Jag är hoppfull och tror att det kan bli skitbra. Men vi måste vara försiktiga. Det här är teknik som är ny. Vi vet inte hur vi ska handska den. Få in kontroll, relevans och tänka på vad man matar in och utbildning. Såna aspekter behöver vi. Sen nån form av konsensus.

Ludwig: Absolut.

Marcus: Att öka själva transparensen då?

Mathias: Ja, exakt. Mina barn, min yngsta, dem tycker fan va bra det är! de fattar inte vad som händer i bakgrunden. Man måste nog verkligen och jag tror på att.. Barn är ju smarta som skjutton att komma och visa där det blivit tokigt asså, för att utbilda dem i detta och att dem kan bli kritiskt tänkande, det är något man lär sig i skolan..Det har blivit tokigt att utbilda dem i detta. De kan bli kritiskt tänkande. Det är nån som har läst i skolan. Framför allt källhänvisning för helskotta. Det här kan ju göra att de blir mer kritiskt tänkande men de måste förstå att det här är text som inte kommer från Wikipedia. Den kanske en blandning av väldigt mycket, men det är nåt som är skapat just då och nu.

Marcus: Tror du att det här öppnar brevet kan skapa någon form av oro som kanske inte är, eller som är befogad för de som inte är vana att använda AI?

Mathias: Är det ett brev som jag borde ha läst?

Marcus: Det är ju det här att de vill pausa…

Mathias: Jag tror att folk… Om man har.. Dem som kanske har, de som alltid ser det värsta i alla sammanhang som tror 'nu så går allt åt helvete', så kommer dem verkligen att bli påverkade av detta mycket mycket mer än en person som har lite mer kunskap inom området. Man skulle kanske vilja nyansera den här debatten och säga att man måste ta paus men tala om varför man behöver det och vad finns det för fördelar och nackdelar i så fall om man gör det hela? Och sen prata om förbud.. Hur bra har dem varit? Vad hände under förbudstiden i Sverige? Man får ju dricka sprit. Det smugglas och ficklas. Det kommer det säkert att göra här med. Men kanske ett mer förtydligande då, inte bara att de smartaste personerna säger att det är farligt. Utan tala om varför tycker de att det är farligt? Vad kan hända? Vad skulle de missa? Få lite mera argument för och emot i så fall. Så att det blir mer begripligt för gemene man i så fall.

Marcus: En vanlig fråga som också kommer upp när man pratar om utvecklingen är att AI kommer att ersätta många mänskliga arbeten.

Mathias: Ja, och det där är nånting som vi hör också väldigt mycket. Hur jävla mycket har inte robotarna i tillverkningsindustrin ersatt mänskligt arbete? Tittar vi hur många bilar det gick åt att bygga en t-ford innan löpande bandet kom till.. Så har det på nåt sätt varit att om vi

ersätter jobb med saker så kommer människorna göra andra saker. Jag är inte så orolig. Det kanske på kort sikt, absolut, kanske sekretera jobb den typen där det
handlar om att producera text och så vidare. Det tror jag att det kanske är. Men jag tror att det kan dyka upp nya jobb som ska träna de här systemen. Att de blir kloka och gör fina och bra saker. På kort perspektiv, absolut. I det längre perspektivet så tror jag att det kommer nog...

Ludwig: Det kommer att ersättas?

Mathias: Jag tror att vi kommer att hitta andra möjligheter när vi interagerar med de smarta maskinerna. Och sen är det vissa saker som jobb som är svårare för en maskin att göra som är mer konstnärligt och kreativitet. Men nu kan man göra konstverk genom att bara prata om att göra det här. Men jag tror att det kanske kan bli en sån här att... Det tycker man häftigt nu. Det är som att.. Ni kanske har LP-skivor - Jag tycker att det är jävla coolt med LP-skivor. Men det börjar komma tillbaka så att det kanske kan bli en sån renaissance. Sen kanske vi blir mätta på det här. Det kanske går i en cykel, vad fan vet jag? Vem hade trott att mina gamla LP-skivor är värda massor? Man kan köpa coola skivspelare. Det var ju helt dött. Så ha kvar videoband kanske är det det som gäller. Det kanske blir att det här... Det är ändå så pass kraftfullt. Men det kanske blir att vi går tillbaka till nånting. Men jag tror att för att avsluta svaren så korta perspektivet så kommer jobb försvinna. Vi har sett att företag blir disruptade för att det kommer teknik som ersätter att de har legat obsoleta. Men människorna har varit bra på att hitta andra saker hela tiden. Vi har utvecklat oss själva, om man bara tar det här med tillverkningsindustrin. Så det behövs andra operatörer och andra typer av tjänster. Vi kanske kommer att utveckla helt nya jobb som vi inte känner till i dag som vi kommer ha behov av. Det kanske kommer behövas mer..
Jag jobbar som... Ja, det har inte med jag att göra. Men jag håller på med som samtalscoach. Där man liksom ska jobba med att få fram potentialen i varje människa och att hitta vad man har för drivkrafter och så vidare. Men man kanske kommer behöva coacha de här jävla robotarna. Att coacha dem på ett sätt, att de blir det schyssta. Så det kanske kommer in helt andra typer av jobb som vi inte tänker på. Men de här smarta systemen kanske behöver en mänsklig coach så att de inte blir fördummade.

Marcus: Det var väldigt stort sett alla frågor som vi hade nedskrivning iallafall.

Ludwig: Jag tycker vi har täckt väldigt mycket av det vi vill få ut av intervjun

Mathias: Ja men jag har inte gjort bort mig då, eller jag?

Ludwig: Nej, nej, verkligen inte

Marcus: Det är ett väldigt svårt ämne.

Mathias: Ja det tycker jag också, Jag kan tycka att ni kan mer än jag. Jag skiter i det här med CHAT eller det gör jag ju inte.. Men vi sysslar inte det på SAS men vi blir berörda av det. Vi jobbar inte direkt med det verktyget. Men det är spännande. Vi har en stor

supportorganisation. Många använder vår mjukvara och någon säger: Vad är det som är fel på min kod? Om man matar in det i chat-GTP då så kan man få fram nånting som i kanske 80 procent av fallen är bra. Men i 20 procent är det felaktigt. Vi pratar om det hela, men när man har matat in saker i systemet- så är det ingen företagshemlighet längre, utan då ägs det av de andra. Så vi har ett regelverk vad vi ska säga och inte säga kring detta, hur vi kommunicerar. Men vi på SAS är inte i de områdena där vi bygger generativa modeller. Vi kan syntetisera data. Här är en datacenter man inte får analysera. Vi kan klona det så att det inte innehåller personer men all den typen av kopplingar, mönster, logik, korrelationer i datacenter. Det område är vi inne i, men vi har ju inte det här att vi... Som kanske våra större konkurrenter lägger in alla sina system, så vi är inte i det. Men däremot pågår det samarbeten att vi tillsammans med Microsoft ska kunna bygga en tjänst som man lägger på deras egen sökmotor- för att på nåt sätt analysera och revidera vad som kommer ut. är det riktigt eller inte då? Vi kan komma in som en tredjepart i det här och kunna lägga på våra...responsable touch på det hela. Så det blir återigen en kringtjänst som kommer att validera resultatet, att det blir schysst hela tiden anpassar man sig.

Marcus: Är det några osäkerheter eller sånt där era kunder brukar komma när ni pratar om nya modeller eller nåt sånt där?

Mathias: Våra kunder på SAS som vi har är läkemedelsindustrin, det kan vara Telekom, det kan vara Bank. När man lånar pengar blir man bedömd av sin återbetalningsförmåga, kreditrisk. Där kan inte ChatGTP gå in. Då använder man historiskt data från kunden. Det ger man inte tillgång till. Många av våra kunder som använder våra lösningar har analyserat sin egen data. Framför allt mycket siffror, historik, transaktioner, interaktioner, kampanjer. Den informationen finns inte tillgänglig i Chat-GTP. Den har tränats på offentlig information. Våra kunders modeller är mer anpassade för ett affärsbeslut, en affärsprocess. Däremot call-center kommer automatisering av sådana så det kommer till ett ärende och där ska kategoriseras kanske till det här och där. Där kan ju liksom det här komma in för våra kunder, men men oftast är våra kunder. De är ju ofta ett större bolag och de har sina krav och så vidare. Men men, visst pratas de om det absolut? Hur man kan kanske gifta ihop det här är någon process och göra anrop med chat-GTP och få fram information. Framför allt kan man skapa content-innehåll. Skriv en catchy slogan eller skapa ett kreativt innehåll. Det kan jag tänka mig att företag kan vilja använda men inte att man stoppar in i en process för att kunna ta fram ett beslut för en kreditgivning eller inte. Många av våra kunder använder vår plattform för att bygga egna modeller för ett visst syfte. Det är inte så mycket att de gör generativa... Det är mycket att generera text, få en fråga och ge text tillbaka.

Marcus: Hur mycket ligger ni på, dem som använder era modeller att det viktigt även när de använder datan att de använder det som ett verktyg i stället för att gå på rå känsla?

Mathias: Det är nånting som jag tror att vi har med oss eftersom vi har varit i det här området sen 1976. SAS har funnits och har det i ryggraden. Det vi kommer ut ska man kunna lita på. I början var det statistiska modeller. Jag vet inte om ni har läst lite statistik. Man kan lätt titta på vad modellen innehåller. Vad är det för koefficienter? Är det relevant?

Det går att göra nån form av rimlighetskontroll. Men när man har kommit in i de mer större maskininlärningsteknikerna där det kan vara ett jättenyrant nätverk som är tränat eller kanske en ensemble av 40 000 besluts är det svårt att gå in och titta på vad som hände. Men då finns det att lägga på tolkningslager på själva resultaten. Det har vi tidigt fått in. När kunder jobbar i vår mjukvara och bygger mer komplexa modeller där det är svårt att titta exakt så har vi det man pratar om, interpretability. Man kan titta på Pd-lies, I chatty värden och såna saker. Och även det som kommer nu, att finns det bias? Finns det nån skevhet i datat från början som modellen lär sig? Det är det man kommer att titta mycket på i AI Act, att man inte diskriminerar. Så vi försöker få in det direkt i vår plattform så att det blir en commodity, det bara ingår. Du behöver inte gå in i något coolt och open-source för att ta fram det. Det är bara att trycka på en knapp. Så kan man få fram. Men det finns en skillnad mellan kön här. Det är inte bra. Då är möjligheten nu med dagens teknik att man kan stävja den här skevheten. Så att vi kan träna om modellen så att skillnaderna i den här variabeln försvinner. Så att det inte finns direkt en skillnad mellan man och kvinna till exempel. Så att teknikerna går framåt så mycket så att just inom det AI Act det är mycket diskriminering. Att man inte ska bli felbehandlad för att man kanske har fel kön eller ras eller ursprung. Eller har någon form av sjukdom. Det kan vi direkt när vi bygger så kan vi få in den kontrollen. Så att för oss är det superviktigt att kommunicera det. Det kan också bli en konkurrensfördel. Det ingår per automatik. Det är inga konstigheter. Jag vill checka den här variabeln för bias. Tryck på en knapp, så är det klart. Det blir enkelt. Förhoppningsvis kommer fler kunder att inse att det är viktigt. Finns det någon skevhet här? Oj, hur ska jag göra det?

9.4 Bilaga 4: Transkribering från intervjun med Mattias Ohlsson

Wictor: Då tänker jag snabbt innan vi börjar med frågor och sådär att dra en snabb… snabb beskrivning av det vi undersöker. Då är det ju osäkerheten kring den stundande AI utvecklingen, vi blev ju då inspirerad av det öppna brevet som även kallats AI-upproret som då är underskrivet av ett flertal AI-forskare runt om i världen. Dom förklarar la främst att vi har tappat kontrollen och att vi behöver pausa utvecklingen för att komma ikapp för att inse vilka potentiella risker som kan inträffa, och där tänker jag… Marcus om du är sugen att ta första frågan där?

Marcus: - Absolut, ja ganska rakt på här då. Hur tänker du kring osäkerheten kring den snabba utvecklingen vi haft nu?

Mattias: För det första, vad menar man med osäkerhet här, låt mig säga såhär. Utvecklingen har ju sen november eller när det nu var gått.. när de släppte Chat-GPT eller va det nu va har ju gått väldigt fort. Man har ju insett att dessa språkmodeller kan göra fantastiska saker och jag har inga aspekter kring det utan jag tycker att det är en spännande utveckling och jag ser inga farhågor i utveckling vad gäller forskning kring det här. Där man kan bli rädd, alla dessa negativa saker som kan inträffa, alltså hur tekniken kan användas på fel sätt. Där kan man bli orolig eftersom det blir lättare och lättare att skapa verktyg som kan skada verksamehter och skada individer och liksom globalt så. Där tycker jag såklart att liksom där är en osäkerhet och eh, men pausa AI utvecklingen som dom vill i den här… det lär ju inte sätta stopp för det typen av utveckling, har jag svårt att se.

Wictor: Men du menar på att den största risken är att det kommer i fel händer då eller?

Mattias: Ja men största, jag ser det som en risk, hur man kan syntetisera tal som gör att man kan lura folk. Att man tror att det är någon annan som ringer, ja men det finns ju hur mycket som helst. Sen hör man ju en del som pratar om att man ska pausa utvecklingen för att man kanske inte vet om AI kommer ta över, om AI kommer göra… att vi får något slags terminator scenario. Det tror jag väldigt lite på. Däremot är det helt uppenbart att det är saker som inte hänger med i den snabba utvecklingen, vi har ju en hel legal sektor som tappar mark här för man vet inte hur man ska hantera dethära. Men så har det väl iförsig alltid varit.

Wictor: Vi har ju varit inne lite på det nu men, i vilka områden ser du att det finns en chans att uppleva osäkerheten. Du var ju inne lite på legala sektorn men har du andra exempel där du känner att det kan uppkomma en viss osäkerhet kring utvecklingen av AI då?

Mattias: Men när ni pratar osäkerhet då menar ni…

Marcus: Alltså att man inte riktigt är säker på konsekvenser och riskerna, vad detta kan leda till då.

Mattias: Mm, nä men som jag sa innan jag tror den största farhågan är alla negativa saker det här kan användas till. Osäkerhet kring, personligen tycker jag det är extremt spännande att försöka pusha och att nå artificiell intelligens, det hade varit super spännande. Jag ser inga risker där, sen har man hört risken med olika kategorier av jobb försvinner men så har det ju alltid varit, när en ny teknik kommit in har en del jobb försvunnits men det har skapats nya jobb.

Marcus: Vad tror du själva utvecklingen kommer leda till eller vad tror du, i en perfekt värld, hur långt kommer man med AI? Kommer det liksom vara autonomt?

Mattias: Det är uppenbart att man tittar på Chat-GPT och GPT-4 och 3, det ser ut om man kan låtas sig imponeras av möjligheterna… man kan nästan tro den kan resonera osv så det ser otroligt avancerat ut när de här, vad modellerna kan göra och det är helt uppenbart att ber du den koda osv så gör den det otroligt bra men jag tror fortfarande inte att denhär GPT-4 är något mer än att du har någonting som har sett jättemycket text. För mig är det fortfarande inget, det finns ingen som knäckt det här att vi kan resonera att bygga upp en modell av världen genom att bara läsa en mängd text. Vi har någon slags transformersmodell som egentligen bara är något sånt här crew fitting modell, jag tror fortfarande att det fattas en viktig komponent för att vi ska nå den här riktigt artificiella intelligensen. Jag tror inte att vi kan gör saker o ting större och större med dom här språk modellerna för att knäcka den. Sen håller jag med om att man måste låta sig imponeras av vad denhär modellen kan göra men jag tror fortfarande att vi inte har knäckt koden än för sann artificiell intelligens. Sen kan man fortfarande som ni uttrycker det, osäker kring vad dethär innebär men jag kan inte överblicka alla negativa eller osäkra… Vad det här kan få till följd med. Det är helt uppenbart att det, det produceras ju en otrolig mängd verktyg på dethär nu så… Jag är inte orolig för domedags tänket men jag är orolig kring att man utnyttjar tekniken i dåligt syfte.

Wictor: Vi funderar lite kring om man kan leva i ett samhälle där AI fungerar helt autonomt, alltså utan stöd eller utan någon mänsklig interaktion, vi refererar till det som “human in the loop”. Men tror du det är en möjlighet, att vi kommer leva i ett samhäller där AI funkar helt autonmt eller om det kommer krävas mänsklig interaktion för dessa verktyg då?

Mattias: En jätte intresant fråga, finns det någonting idag som fungerar helt autonomt? Det finns om vi tar de börsen, jag vet inte men det är väl ett exempel på att det finns otroligt mycket verktyg som styr börshandel idag som, undra hur mycket det är av det som är autonomt och vad som styrs av algoritmer som ett exempel. Vi strävar la att få bilar som är självkörande och det skulle ju då i en väldigt begränsad miljö där vi har någonting som är helt autonomt. Men jag funderar på hur, finns det sånna miljöer idag där vi kan säga saker och ting lever mer eller mindre av sig självt utan att vi har för mycket mänsklig kontakt. Därför att vi har valt att göra på det sättet. Att vi vill ha system som sköter sig mer eller mindre autonomt utan att vi blandar oss i för mycket, man kan säkert hitta sånna små subsystem som redan finns idag så frågan är vad händer om vi skalar upp det här till större och större saker

Marcus: Vad heter det, dem här OpenAI, dem som har utvecklat ChatGPT där, vissa där har ju sagt det att dem känner att det går lite för fort och att dem inte vet exakt vad det är som sker eller hur den beräknar fram resultat och sånt där. Vilka problem eller risker tror du att det kan skapa om man fortsätter utvecklingen i samma takt utan att tänka på någon typ av reglering.

Mattias: Det är svårt… Det är en jättesvår fråga tycker jag… När dem säger att dem inte riktigt vet hur den kommer fram till alla sina svar så är det… På något sätt så hintar dem att i dem här modellerna skapat någon form av så komplex modell så den börjar bete sig som om den vore intelligent.

Marcus: Det är väl lite hur dem räknar fram, säg att man skriver in någonting och så beter sig chatboten lite bias eller något sånt där och då vet dem inte varför den har blivit så på den datan som de fört in, alltså hur den räknar fram resultatet.

Mattias: Jag menar, det är ju ingen mystik bakom detta, det är ju modellen i sig som gör detta ganska enkelt att penetrera, liksom att man går in i modellen och försöker förstå hur den genererar sina svar så är det månt och mycket sannolikheten. Sannolikheten att generera nästa ord och det är princip det som är då, kan vara svårt här förstås att vi som en enskild individ har inte den möjligheten att se så mycket textmassa som dessa modellerna har sett, vi kan liksom inte överblicka allt det. Och bygger man in bias i det här så kommer ju naturligtvis den biasen, jag tror inte att det är något annat utan biasen kommer ju från all den text som den har sett. Jag tror inte det händer något mystiskt eller någonting eller autonomt, det har jag svårt att se, för mig är det liksom endast fortfarande en konsekvens av det den har sett, otroligt otroligt mycket text.

Marcus: Det har alltså med den insamlade datan att göra då? Den drar den slutsatserna från den datan? Helt enkelt?

Mattias: Ah, precis. Det är alltså en konsekvens av all den textmassan modellen har sett att den då kan bli bias. Jag kan inte se någon annan mekanism än det och den är ju välkänd oavsett om man inte jobbar med dem här stora modellerna utan när man jobbar med maskininlärning och gör prediktioner osv man tränar den på data och har du bias i din data så kommer du att se den, då kommer din modell bli bias. Jag har svårt att tänka mig att det finns någon som har någon typ av överblick över den textmassan som man matar in i modellerna, det är liksom väldigt mycket.

Marcus: Då tänker jag lite spontant, tror du att man kommer kunna komma runt det här på något sätt? att man får in data som är helt neutral och korrekt.

Mattias: Ja det är en jättesvår fråga. Hur ska man kunna… Hur ska man kunna styra det här så man får något som är helt neutralt och vem ska bestämma vad som är neutralt, det blir jättesvårt. Jag menar det finns väl någon slags tro här att textmassan den har sett är mer västvärldsinriktad mer än vad den här södrafrikainriktad. Vilket innebär att den kommer ha en

bias mot västvärlden snarare än afrika. Det är väl helt rimligt. Och vem avgör vad som är normalt? Och då kommer man till det här problemet med reglering att det går så väldigt fort. Behöver det finnas någon form av legat instrument som talar om vad är det för korrekta textmassor du får använda och vad får du inte använda osv.? Där hänger man ju inte med uppenbarligen. Det i sig kan ju ge upphov till en av de här osäkerheterna.

Wictor: Jag tänker då kan man.. skulle man kunna lita på resultatet AI genererar om man inte förstår hur datan är insamlad, vilka regler som är tillämpande under utvecklingen

Mattias: Nä, det är ju en superutmaning. Hur kan man lita på de här sakerna den genererar. Det är jättesvårt. Det ända du kan lita på är om du har extremt god kunskap om området själv.

Wictor: Nä, för jag var lite nyfiken på den här Chat-GPT, så jag testade lite kluriga frågor för att se om den svarade rätt eller fel. Den svarar gärna väldigt självsäkert gentemot mig som användare men det känns mer som en typ av sannolikhet att det kan vara så här att.. men i vissa fall var det helt fel.

Mattias: Sen spelar det väl roll om du använder gpt-4 eller gpt-3. GPT-4 har väl blivit snäppet vassare men absolut det finns väl fortfarande vissa frågor man ställer den som man tänker att den här kan uppenbarligen inte resonera.

Marcus: Finns det något sätt man kan… För nu tänker jag som en användare som inte är jättevan vid att använda AI, som man är medveten om i alla fall, finns det något sätt som forskarna eller utvecklarna kan ge mer förtroende för AI om man kan säga så? kan man minska riskerna för misinformation

Mattias: Det hade varit spännande och då tror jag och jag ser fler. Det handlar om, gissar jag, då att det handlar om någon slags kontroll på vad du använder för textmassa när du tränar den, det är ju ett sätt. Sen är den andra svårigheten då att fakta är fakta så det går ju inte.. där finns det ju en sanning men man kan ju många gånger be om att få åsikter om saker och ting. och åsikter bygger på uppfattningar och där har de genast en massa bias, så det blir ju jättesvårt, jag har inget bra svar där.

Marcus: Det öppna brevet då som föreslår att man ska ta en 6 månaders paus, eller minst 6 månader. tror du att det är nödvändigt på något sätt att man pausar för att förhindra att det ska bli några större konsekvenser eller tror du att utvecklingen liksom den går som den går som man säger.

Mattias: Asså.. För att någonting sådant ska bli effektivt, då säger jag att man skulle halta all utveckling av stora språkmodeller i 1 år för att man behöver sätta sig ner, diskutera, förstå och liksom.. För att det ska vara effektivt behöver alla göra det. Jag tror det är helt omöjligt att få alla att göra det. Framförallt inte dem som har onda avsikter, kommer naturligtvis inte halta en sån utveckling. Så även om tanken är god och jag kan förstå mycket av resonemangen, förstå mycket av oron just, men om inte alla gör detta så blir det svårt. Och

när en sån person som elon musk gör detta så har jag svårt att tro honom för ibland undrar jag om han.. jag menar att han har ju ett kommersiellt syfte med sitt egna företag, så han kanske bara vill pausa för hans egna företag ska komma ifatt, så det är jag inte så mycket för, då är det väl mer andra, mer neutrala forskare som bland annat Hinton(?) som börjar också säga det här då kanske man ska ta det mer på allvar. Men Elon Musk har jag lite svårt att tro på, i det här avseendet iallafall.

Wictor: Vi hade en följdfråga där om det ens är möjligt att stoppa utveckling.

Mattias: Jag tror det i praktiken är otroligt svårt. Jag menar hur ska man kontrollera detta? Det går ju liksom inte.

Wictor: Om vi går vidare då, vi snackade lite om konsekvenserna innan också men, vad ser du som dem största och kanske möjligaste konsekvenserana av AI utvecklingen.

Mattias: Det är väl helt upp till.. det finns ju ett antal positiva konsekvenser som jag kan se. Det är ju idag väldigt lätt att få en.. Att det finns massa verktyg för automatisering som bygger på dem här sakerna som uppenbarligen gör att saker och ting är effektivare och går snabbare. Så där finns ju en bra konsekvens av.. men sen är det ju otroligt svårt, jag har svårt att överblicka alla möjliga effekter och alla möjliga användningsområden om man skalar upp det här.. Den frågan tycker jag är jättesvår. Tiden kommer ju att utvisa.

Marcus: Jag tänker all den här media uppmärksamheten dem har fått nu, Chat-GPT och allting. Tror du att det kan ha potentiellt skadat utvecklingen av AI på något sätt, att det liksom bygger upp en slags rädsla som kanske inte är befogad.

Mattias: Jo men så kan det vara. Vi använder ju AI-tekniker idag hela tiden som vi inte känner till och det är mycket som är gott för mänskligheten. Men då när den får den här uppmärksamheten som då har en negativ klang på sig naturligtvis så kan de ju få, det behöver ju inte alltid, det kan leda till för mig överdriven negativitet kring AI-området absolut.

Marcus: Tycker du att man behöver vara orolig att AI kommer ta över arbeten?

Mattias: Jo, men så har det väl alltid varit med nya tekniker, nya tekniker leder till förändringar inom hur vi jobbar. När datorer kom var det många, jag menar nu förlorar vi alla jobb, datorerna tar över allting så är det naturligtvis inte. Utan saker och ting förändras hela tiden och nya jobb skapas. Idag kan du bli anställd som patentingenjör kanske inte i sverige men i USA iallafall. Nya jobb skapas, gamla jobb försvinner, så har det nog alltid varit. Jag kanske är överdrivet positiv här men jag tror inte att det här med just AI och ta över alla jobb är något speciellt.

Wictor: Det kommer rättare sagt balansera upp sig, nya jobb kommer och gamla försvinner?

Mattias: Ja så har det väl egentligen alltid varit i historien. Jag tror säkert att detta teknikskiftet kommer få motsvarande effekter.

Wictor: Det var alla frågor vi hade, tack så jättemycket.

Mattias: Tack, ni får gärna skicka arbetet till mig sen om ni vill, kan vara roligt att läsa.

9.5 Bilaga 5: Intervjuunderlag

9.5.1 Intervjuförfrågan

Hej,

Jag, Marcus Malmberg och Ludwig Carlsson studerar sista terminen inom informatik och är nu i en fas där vi letar efter möjliga respondenter till vår kandidatuppsats där vi undersöker osäkerheter kring AI-utvecklingen (inspirerat av AI upproret från Elon Musk mm). De områden som undersökningen kommer behandla är osäkerheten med AI-utvecklingen, om osäkerheten är befogad eller inte samt riskerna/problematiken med att låta utvecklingen av AI fortsätta utan människans förståelse och kontroll.

Vi undrar därför om du är intresserad att svara på ett antal frågor antingen via Zoom/Teams eller fysiskt om det är möjligt. Intervjun kommer ta max 1 timme och skulle hjälpt oss enormt för att genomföra vår forskningsstudie.

Vänliga hälsningar,
Wictor, Marcus och Ludwig

9.5.2 Introduktion till intervju

Hej! Vi tänker att vi drar igång direkt och vill innan intervjun börja först med att tacka dig för att du tar dig tiden och hjälper oss genomföra denna forskningsstudie. Vi vill också dubbelkolla ifall det går bra om vi spelar in hela intervjun och om vi får nämna dig i rapporten eller om du vill vara anonym? Inspelningen är då endast för att vi ska transkribera intervjun efteråt och det gör det enklare för vårt arbete.

Startar inspelning

Vi undersöker då osäkerheten kring den stundande AI utvecklingen och blev inspirerade av det öppna brevet, även kallat AI upproret som är underskrivet av bl.a Elon Musk och Steve Wozniak. De förklarar främst att vi människor har tappat kontrollen och behöver pausa utvecklingen för att komma ikapp och förstå vilka potentiella risker och konsekvenser som kan inträffa.

9.5.3 Obligatoriska intervjufrågor

Osäkerheten kring den stundande AI-utvecklingen

- Vad är dina tankar om osäkerheten kring AI-utvecklingen?
- Anser du att vi behöver vara oroliga över dagens AI-utveckling?
- (Följdfråga) Inom vilka områden inom AI anser du det finns störst chans till att uppleva osäkerhet?
- Anser du att AI-utvecklingen är utom kontroll? Varför anser du detta?
- Kan vi leva i ett samhälle där AI fungerar autonomt eller bör det alltid finnas en human in the loop? kommer vi leva i ett
- Hur hade man kunnat utveckla AI-systemen för att öka säkerheten och förtroendet?

Risker/problematik

- Vilken typ av risker/problematik finns det med att fortsätta utvecklingen av AI utan mänsklig förståelse och kontroll?
- Hur ska man lita på resultatet AI genererar utan att förstå hur datan är insamlad eller vilka AI regler som är tillämpade under utvecklingen?

Konsekvenser

- Tycker du att en utvecklingspaus på minst 6 månader hade varit nödvändigt eller tror du man kan låta AI utvecklas som den gör idag utan några större konsekvenser? Vilka konsekvenser hade annars kunnat uppstå?
- Vad tror du hade kunnat vara den värsta tänkbara konsekvensen av dagens AI-utvecklingen?
- Bör människor vara oroliga över att AI kommer byta ut mänskliga arbeten?
- (Följdfråga)Tror du de jobb som försvinner möjligen balanseras upp med nya "AI-jobb" eller kommer arbetslösheten generellt sätt öka?

9.5.4 Intervjufrågor utöver det obligatoriska underlaget

Övriga intervjufrågor

- Är AI utvecklingen mot en emotionell och social förståelse rimlig? Kommer risken för kränkningar eller olyckliga situationer alltid finnas?
- Vem bär ansvaret om resultatet är felaktigt, kränkande eller missvisande? Ligger ansvaret på användaren, utvecklaren? Kanske situantionanpassat (tänker på teslabilen som kraschade)
- I DI framgår det att ni främst arbetar med "Skapa medvetenhet och vägleda organisationer i hur de på bästa sätt kan implementera AI på ett ansvarsfullt sätt" Vad är de främsta anledningar att organisationer avstår eller kan vara motstridiga mot implementering?
- Tycker du AI's uppmärksamhet i dagens media påverkar organisationers syn på att implementera det? Skapar de en oro som är eller inte är befogad?
- Hur hög tillförlitlighet känner du med den utdata AI genererar. Kan man lita på den data?

Besöksadress: Allégatan 1 · Postadress: 501 90 Borås · Tfn: 033-435 40 00 · E-post: registrator@hb.se · Webb: www.hb.se