Strål säkerhets myndigheten
Swedish Radiation Safety Authority

Författare: Martin Castor
Jonathan Borgvall

Geistt AB

2019:29

Myndighetsstöd: Artificiell Intelligens – tillämpning inom kärnkraften

Rapportnummer: 2019:29 **ISSN:** 2000-0456
Tillgänglig på www.stralsakerhetsmyndigheten.se

SSM perspektiv

Bakgrund

Området artificiell intelligens (AI) genomgår en omfattande utveckling och stora satsningar görs inom FoU på olika håll runt om i världen. Vinnova fick under 2017 i uppdrag av regeringen att genomföra en kartläggning och analys av hur väl artificiell intelligens (AI) och maskininlärning kommer till användning i svensk industri, offentlig sektor och det svenska samhället samt vilken potential som kan realiseras genom att stärka användningen. När det gäller AI som relaterar till kärnkraften finns internationella exempel med praktiska tillämpningar och användningsområden. Mot denna bakgrund initierades detta myndighetsstöd som en del i den omvärldsbevakning som görs. Svensk kärnkraft har redan genomgått en omfattande moderniseringsfas men myndigheten behöver inhämta mer kunskap om AI generellt eftersom denna begreppsvärld är tämligen otydlig. En annan aspekt är om AI används i någon form eller det planeras för detta inom svensk kärnkraft. Målsättningen var i första hand att:

a) få en orientering kring definitioner av de olika begrepp vilka är sammanlänkade med AI som portalbenämning för ”området”

b) genomföra en kartläggning med aktuell bild över hur utvecklingen ser ut med de planer eller tillämpningar som i första hand finns inom svensk kärnkraftindustri med syfte och mål för tillämpningarna.

Resultat

Rapporten ger en ganska komplex men överskådlig bild med definition av de begrepp som relaterar till AI och konstaterar att användningen inom kärnkraftindustrin är mycket liten. Däremot är tillämpningen av AI på frammarsch inom ett annat av myndighetens tillsynsområden nämligen sjukvårdssektorn.

Relevans

Rapporten har skapat ett internt intresse och kan fungera som utgångspunkt för andra projekt med en mer avgränsad inriktning exempelvis mot tillämpningar inom sjukvårdssektorn och dess eventuella strålsäkerhetsrisker.

Behov av vidare forskning

Det finns inga direkta behov av ytterligare myndighetsstöd med denna inriktning men kan fungera som underlag för annan forskning med inriktning mot tillämpningar inom sjukvårdssektorn.

Projekt information

Kontaktperson SSM: Steve Selmer KM
Referens: SSM2018-5852

Strål
säkerhets
myndigheten
Swedish Radiation Safety Authority

Författare: Martin Castor
Jonathan Borgvall

Geistt AB

2019:29

Myndighetsstöd: Artificiell Intelligens – tillämpning inom kärnkraften

Datum: December 2019
Rapportnummer: 2019:29 **ISSN:** 2000-0456
Tillgänglig på www.stralsakerhetsmyndigheten.se

Denna rapport har tagits fram på uppdrag av Strålsäkerhetsmyndigheten, SSM. De slutsatser och synpunkter som presenteras i rapporten är författarens/författarnas och överensstämmer inte nödvändigtvis med SSM:s.

Myndighetsstöd: Artificiell Intelligens – tillämpning inom kärnkraften

Martin Castor & Jonathan Borgvall, GEISTT AB

Sammanfattning

Rapporten utgör en del av SSM:s kunskapsuppbyggnad rörande AI, Artificiell Intelligens, genomförd inom ramen för ett myndighetsstödjande uppdrag till GEISTT AB under våren 2019. Rapporten förklarar på en övergripande nivå AI-området och implikationer för SSM:s verksamhet, med specifikt fokus på tillämpningar avseende operativ kärnkraftssäkerhet enligt överenskommelse med SSM.

Innehåll

1 Bakgrund 5
1.1 Syfte 5
1.2 Avgränsningar 6
1.3 Läsanvisningar 6
2 Vad är AI? 8
2.1 Olika målsättningar 11
2.2 Vilken typ av AI är det? 15
2.2.1 Maskininlärning 15
2.2.2 Symboliska och subsymboliska ansatser 15
2.3 Särdrag 21
2.4 Nätverksstruktur 21
2.5 Träningsmetoder 24
2.5.1 Övervakad inlärning 24
2.5.2 Oövervakad inlärning 25
2.5.3 Förstärkningsinlärning 26
2.6 Djupinlärning 27
2.7 Andra ”heta begrepp” 28
2.7.1 Transfer learning 29
2.7.2 GAN 29
3 AI-tillämpningar 31
3.1 Tillämpning per funktion 33
3.2 Exempel på tillämpningar 34
4 Humancentrerad automation 37
5 Organisatoriska, etiska och juridiska aspekter på AI 45
5.1 EU JRC perspektiv 46
5.2 EU HLEG AI perspektiv 47
5.3 Asilomars AI-principer 57
5.4 AlgoAware perspektiv 58
5.5 Kondensat av etiska principer 59
5.6 Utvecklingsprocess 59
6 Aktuella satsningar 60
6.1 Amerikanska forskningsprojekt 60
6.1.1 MEITNER 60
6.1.2 I4Gen 60

6.1.3 Explainable AI (XAI) 60
6.2 Nationella AI agendor inom EU 62
6.3 EU finansierade projekt 62
6.3.1 RAIN 62
6.3.2 AI4EU 62
6.3.3 AlgoAware 63
6.3.4 European AI alliance 63
6.3.5 CLAIRE 63
6.3.6 SIENNA 63
6.3.7 SHERPA 63
6.3.8 PANELFIT 63
6.4 Nordiska AI-satsningar 63
6.4.1 Uniper - OKG 63
6.4.2 WASP 64
6.4.3 MonitorX 64
6.4.4 CHAIR 64
6.4.5 AI Innovation of Sweden 64
6.4.6 SAIS 65
6.5 Internationella företags produkter och arkitekturer 65
6.6 ISO 66
7 Framåtblick 68
8 Referenser 74

Nyckelord:

AI, Artificiell intelligens, maskininlärning, neurala nätverk, djupinlärning, beslutsstöd, säkerhet

1 Bakgrund

Syftet med denna rapport är att ge en bred översikt över området artificiell intelligens (AI) för att stödja SSM:s handläggare och deras kunskapsuppbyggnad inom området. Rapporten har sammanställts av GEISTT AB som ett myndighetsstödjande uppdrag för SSM under våren 2019. Uppdraget har genomförts med Steve Selmer från SSM:s enhet för Människa-Teknik-Organisation (MTO) på avdelningen för kärnkraftssäkerhet som projektansvarig, men beskrivningen av AI som område är mer generell och bedöms vara användbar för flera enheter på SSM utöver MTO-enheten.

AI-området är stort, diversifierat och relativt svårdefinierat, vilket kommer utvecklas vidare i rapporten. Som både Regeringskansliet (2018) och Vinnova (2018) påpekar så finns det inga entydiga eller allmänt vedertagna definitioner av AI, och individers uppfattning påverkas av också av media och populärkultur. Ett av rapportens syften är därför att beskriva viktiga skiljelinjer och olika perspektiv på AI samt att ”avmystifiera” området.

De senaste åren har viktiga tekniska genombrott skett som gör att produkter som använder sig av olika typer av algoritmer från AI-området har ökat markant. AI som forskningsområde har dock förekommit sedan 1950-talet. Diverse tekniker och algoritmer som utvecklats inom AI-området används redan idag i relativt hög utsträckning i vardagligt tillgängliga produkter som exempelvis skräppost-filter, bildigenkänning och taligenkänning. Utsträckningen i användning och antalet tillämpningar bedöms dock komma att öka markant.

Relevansen för SSM och de kärntekniska tillståndshavarna är, och framför allt kommer att bli, hög allt eftersom tillämpningarna mognar. Digitaliseringen av kärnkraften ökar, både av effektivitetsskäl och eftersom det i vissa fall är svårt att få tag på analoga reservdelar till kärnkraftverken. Både tillverkare och kärnkraftverk ökar därför sina satsningar på digitalisering och s.k. *wireless* teknik. Begreppet AI representerar ett brett fält av forskning och utveckling, med många olika typer av beprövade och kommande tillämpningar, men det finns ingen global konsensus på hur AI-baserade system och tekniker kan eller bör tillämpas. Rapporten syftar till att bidra till ökad förståelse hos hos SSM och tillståndshavarna rörande AI, så att kommande satsningar och system kan analyseras med högre precision, samt att ge förståelse för hur AI-baserade system kan ingå i större sammanhang med bibehållen eller ökad säkerhet och effektivitet.

Det är inte meningsfullt att reproducera tidigare historiska översikter gällande AI-områdets framväxt. För den mer detaljerade historien bakom AI-områdets framväxt hänvisas till referensverken *Artificial Intelligence – A Modern Approach* (Russel & Norvig, 2010)[1] och *Deep Learning* (Goodfellow, Bengio, & Courville, 2016)[2]. Båda dessa böcker används i hög utsträckning som grundlitteratur vid AI-kurser världen över, och används i denna rapport som utgångspunkt för den övergripande beskrivningen av området.

1.1 Syfte

Syftet med det myndighetsstöd som redovisas i rapporten är att kartlägga tillämpningen av AI, primärt inom svensk kärnkraftindustri.

Målsättningen för arbetet har varit att:

1. få en kortfattad orientering kring definitioner av centrala begrepp som används inom AI-området

1 http://aima.cs.berkeley.edu
2 http://www.deeplearningbook.org

2. genomföra en kartläggning som ger en aktuell bild över AI-utvecklingen och AI-tillämpningar som i första hand finns inom svensk kärnkraftindustri
3. få en sammanfattande beskrivning av de internationella trender som kan identifieras
4. få en allmän översikt med exempel på potentiella risker som kan vara förenade med tillämpningen av AI inom kärnkraftsindustrin

1.2 Avgränsningar

Rapporten är utformad för att ge en översikt över det breda AI-området, utan att gå alltför djupt ner i tekniska detaljer. Närmast forskningsfronten finns en mycket stor mängd olika begrepp och varianter av olika ansatser. Denna rapport beskriver dock endast grövre skiljelinjer.

Satsningar på AI görs idag på stor bredd inom olika branscher med många olika tillämpningar. I rapporten nämns några specifika satsningar som relaterar till AI inom kärnkraft, och med fokus på drift av kärnkraftverk. Tillämpningar som rör exempelvis generellt planeringsarbete eller inpassagekontroll/övervakning berörs inte, även om sådana skulle kunna ha relevans vid kärnkraftsanläggningar. Exempel på sådana generella tillämpningar som inte berörs kan vara AI-algoritmer för skräppostfilter till e-post, personalplanering, ansiktsigenkänning för inloggning/inpassering, och liknande användbara, men mer generella tillämpningar. Den utveckling och användning av AI som berör medicinteknisk verksamhet, inklusive alla AI-tillämpningar som finns där, har också avgränsats bort i denna rapport, eftersom utvecklingen inom detta område eventuellt kommer beskrivas i en separat översikt.

Enligt överenskommelse med SSM så fokuserar denna rapport på specifikt på tillämpningar avseende operativ kärnkraftssäkerhet.

1.3 Läsanvisningar

Rapporten består av sex huvudsakliga avsnitt:

- **Vad är AI?** Detta avsnitt är avsett att ge grundläggande förståelse för vad AI är. Avsnittet börjar med en teoretiskt inriktad del där en översikt av området presenteras. Relationen till ett antal andra områden beskrivs kortfattat. Därefter presenteras en fördjupning gällande så kallad djupinlärning *(deep learning)*, eftersom detta är det för tillfället mest framgångsrika och omtalade delområdet inom AI.
- **Exempel på tillämpningar:** Detta avsnitt beskriver ett antal exempel på AI-tillämpningar, både inom och utanför kärnkraftsdomänen.
- **Automationsfrågeställningar:** Många av de AI-relaterade frågor som är aktuella idag har diskuterats och analyserats under lång tid inom den humancentrerade automationsforskningen, som av SSM brukar hanteras inom området MTO (Människa-Teknik-Organisation). I detta avsnitt återges exempel på teoretiska modeller och rekommendationer från detta forskningsområde.
- **Aktuella projekt och satsningar:** I detta avsnitt beskrivs ett antal större AI-relaterade projekt och satsningar som bedöms vara relevanta för SSM kunskapsuppbyggnad. Delar av detta stycke är resultatet av ett antal intervjuer med AI-intressenter inom svensk energiproduktion som genomförts under våren 2019.
- **Organisatoriska, etiska och juridiska aspekter på AI:** Detta avsnitt beskriver ett antal organisatoriska, etiska och juridiska aspekter på AI som bedöms vara intressanta för SSM. Flera av de sammanställningar som återges här skulle efter gallring av innehållet kunna vara användbara som utgångspunkter om SSM någon gång i framtiden får uppdrag att genomföra tillsyn eller författa föreskrifter som rör AI-baserade system.

- **Framåtblick**: Detta avsnitt innehåller information avseende den förväntade framtiden för AI-området.

Rapporten är avsedd att vara möjlig att läsa både snabbt, för att få en grundförståelse, och långsamt som inledning till fördjupad förståelse. Exempelvis, under det första avsnittet som beskriver vad AI är så ges en mängd referenser och länkar till externa resurser där filmer och demos finns tillgängliga på Internet. För djupare förståelse av flera av de mer abstrakta koncept och begrepp som används inom AI-området så rekommenderas läsaren att följa de länkar som anges.

På motsvarande sätt är avsnittet kring organisatoriska, etiska, och juridiska aspekter avsett som ett avsnitt som både kan läsas översiktligt, och även som referensbas om SSM får i uppgift att ta fram riktlinjer rörande AI.

Ovanstående upplägg och syfte innebär även att i denna rapport, som i sig är skriven på svenska, återges en hel del exempel och uppräkningar på engelska. Syftet med detta är dels att inte förlora information i samband med översättning, dels att ge läsaren de engelska begreppen i olika sammanhang, för att underlätta fortsatt inläsning i de olika länkar och referenser som ges.

2 Vad är AI?

Artificiell intelligens (AI) är ett i nuläget mycket snabbt expanderande område, både som forskningsområde och som tillämpad teknologi i allt fler produkter. Som forskningsområde är det dock inte nytt utan brukar sägas ha sin start 1956 vid en nu berömd forskningskonferens på Dartmouth College i New Hampshire, vid vilken uttrycket sägs ha myntats första gången.

AI bör ses som ett samlingsnamn för en större familj av delvis likartade, delvis relativt olika tekniker för att utveckla datorprogram med problemlösningsförmåga. Som Russel och Norvig (2010) påpekar skulle troligen begreppet beräkningsbar rationalitet *(computational rationality)* varit ett lämpligare begrepp för att förstå innebörden. Det handlar alltså om att utveckla datorprogram som kan komma fram till slutsatser som uppfattas som rationella, ofta genom att kunna urskilja mönster i någon datamängd och sedan agera på ett relevant sätt utifrån detta.

Någon etablerad och ensad definition av vad AI är finns inte idag (EU JRC, 2018), men det finns naturligtvis en uppsjö förslag. EU High level expert group on AI (EU HLEG AI), som samlat ett större antal AI-experter, publicerade nyligen dessa två definitioner:

> *Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications). (EU HLEG AI, 2018a).*

I en uppdaterad definition används följande formulering:

> *Artificial intelligence (AI) refers to systems designed by humans that, given a complex goal, act in the physical or digital world by perceiving their environment, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best action(s) to take (according to pre-defined parameters) to achieve the given goal. AI systems can also be designed to learn to adapt their behaviour by analysing how the environment is affected by their previous actions. As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems). (EU HLEG AI, 2018b).*

Överblicken över området försvåras av flera faktorer:

- AI-utvecklingen kännetecknas av att den inbegriper många varianter inom familjer av beräkningstekniker, algoritmer, och problemlösningsförmågor, som ofta har egna namn. Alla dessa kallas i denna rapport för AI-ansatser. Nya begrepp och sätt att namnge sin AI-ansats tas fram med hög hastighet nära forskningsfronten.
- Dessa olika AI-ansatser används för många olika tillämpningar, där popularitet och entusiasmen för de olika ansatserna förändras över tid.
- Utvecklingen av AI, både inom forskningsområdet och för produkttillämpningar, går för tillfället mycket snabbt och AI är ett mycket omnämnt *buzzword*.

- Olika falanger av forskare inom AI-området förespråkar olika synsätt och AI-ansatser. När man kommer ner i detaljerna och nära forskningsfronten finns det en mängd olika begrepp och varianter av AI-ansatser som kräver djup specialistkunskap för att kunna särskilja från varandra. Det är heller inte alltid kommunikationen mellan dessa olika falanger är fullt fungerande och de kan också i viss utsträckning konkurrera om samma forskningsmedel. En beskrivning av olikheterna finns hos Domingos (2015) som beskriver AI-forskare som tillhörandes fem huvudsakliga ”stammar”, beroende på vilka AI-ansatser de föredrar:
    - *Symbolists use logical reasoning based on symbols* (använder t.ex. expertsystem).
    - *Connectionists build structures inspired by the human brain* (använder t.ex. artificiella neurala nät).
    - *Evolutionaries use methods inspired by Darwinian evolution* (använder t.ex. genetiska algoritmer).
    - *Bayesians use probabilistic inference* (använder t.ex. bayesianska nätverk).
    - *Analogizers extrapolate from similar cases seen previously* (använder div. statistiska ansatser).
- Åsikten vad som bör räknas om AI förändras också över tid, allt eftersom lösningar blir överkomliga och kända, vilket ibland refereras till som AI-effekten. Ett exempel kan vara igenkänning av naturligt talat språk vilket av många ansågs vara AI för tjugo år sedan, medan vissa nu anser att det bara är ”språkbehandling”. På samma sätt kan man diskutera om en schackdator är intelligent. Sökalgoritmer som används för att söka igenom stora tillståndsrymder, som i en schackdator, utgjorde länge en stor del av AI-forskningen, men dessa sökalgoritmer lärs nu ut på grundkurser för datavetare. Det finns ett gammal skämt som säger att AI är “coola saker som som datorer inte kan göra”, vilket innebär att så snart någon löst problemet så anses det inte riktigt vara AI längre. Positionen på andra änden av skalan uppvisas av de, ofta försäljare, som beskriver all någorlunda fyndig eller smart kod som AI.
- Moderna tillämpningar använder sig ibland av hybrida tekniker där man kombinerar förmågor från flera AI-ansatser.
- Vad som innefattas i begreppet AI är inte standardiserat och synsätten är flera. Maskininlärning anses exempelvis av vissa vara synonomt med AI, av andra inte, även om inlärningsförmåga i koden dock kan anses vara en av de viktigare skiljelinjerna. Med maskininlärningsförmåga avses här att systemet efter en viss träningsperiod kan göra något som programmeraren inte programmerat in från början. För vissa är det alltså AI så fort det är en någorlunda ”smart lösning”, medan andra ställer självlärande och proaktivitet som krav för att kalla det AI.
- Att andra *buzzwords* ofta nämns i samband med AI gör också överblicken svårare. Exempel på aktuella sådana begrepp är:
    - ***VR, Virtual Reality:*** En datorgenererad värld visas för en mänsklig användare genom en huvudburen presentationsyta/VR-headset.
    - ***AR, Augmented Reality:*** Datorgenererad information presenteras överlagrat på en vy av verkligheten.
    - ***MR, Mixed Reality:*** En datorgenererad värld där (vissa) virtuella objekt även har en fysisk motsvarighet som användare kan interagera med.
    - ***Connectivity/wireless:*** Processorer och applikationer är uppkopplade mot någon form av nätverk och har därför förmågan att utbyta data.
    - ***Internet of Things (IoT):*** Många mindre och enklare processorer och ”prylar” som kan samla data och styras digitalt.
    - ***Big data:*** Insamling, tillgång till, och analys av mycket stora datamängder.

- **Automation/autonomitet:** System som har förmågan att agera självständigt. Oftast är de dock inte helt autonoma utan mänsklig styrning av systemen finns på någon nivå.
- **Robotik:** Robotik avser i denna rapport den forskning, utveckling och operativ användning av system som har en fysisk befintlighet eller manipulationsförmåga på den fysiska världen. Den fysiska delen av roboten styrs ofta av ett system som använder sig av diverse AI-tekniker.
- **Intelligent/digital assistent eller agent:** I denna rapport avses system baserade på diverse AI-tekniker som inte har en befintlighet i den fysiska världen, utan uteslutande finns i den digitala världen.
- **Datorseende/maskinperception:** Utveckling av datorers förmåga att exempelvis se och höra.
- **Digital tvilling:** Anläggning eller system som har en kontinuerligt exekverande spegling/simulering av sina processer vilket möjliggör prediktioner, analys av omfall, underhållsplanering, och felsökning med hjälp av den digitala tvillingen.
- **Digitalisering:** Överordnat begrepp för många av begreppen ovan, där det sker en övergång från analoga till digitala tekniker exempelvis för processtyrning.

Observera att en produkt som använder sig av diverse AI-ansatser ofta är komponenter i begreppen ovan och kan utgöra en nödvändighet eller förstärkning för varandra i något specifikt tillämpningsfall. Exempelvis behöver kanske en produkt som använder sig av någon AI-ansats, som exempelvis *deep reinforcement learning*, ha kontinuerlig tillgång till mängder av data (*big data*) från många enklare sensorer (*IoT*), som den sedan visas för en mänsklig användare genom en VR-hjälm (*VR*), i syfte att erbjuda en digital tvilling som ger människan förmågan att bättre förstå underhållsbehoven i en automatiserad produktionsanläggning.

AI bör betraktas som ett paraplybegrepp som spänner över diverse beräkningsansatser som exemplifieras i listan nedan och som används i många olika tillämpningar. Listan är dock långt från komplett och det finns hundratals, om inte tusentals, olika typer av algoritmer som används inom AI-området. För exempel på tillämpningar se avsnitt 3.2.

- Bayesianska nätverk
- Dolda markov modeller (*Hidden Markov Models, HMM)*
- *Constraint based reasoning (CBR)*
- Genetisk programmering
- Evolutionära algoritmer
- Expertsystem
- Intelligenta agenter
- Beteendeträd
- Naturlig språkbehandling *(Natural Language Processing, NLP)*
- Artificiella neurala nätverk
- Djupinlärning

AI med delområdet maskininlärning är vetenskapliga områden som bör betraktas som delar av datavetenskapen, även om det finns starka relationer till flera andra vetenskapliga områden som exempelvis statistisk. AI kan i många fall sägas vara tillämpad statistik. AI är alltså en samling koncept, problem/uppgifter man vill lösa, och metoder för att lösa problemen/uppgifterna. Det är ett brett begrepp och diverse algoritmer som vuxit fram inom olika AI-ansatser kan användas till i princip ”vad som helst”, givet att man har ett lämpligt beräkningsproblem att lösa, vilket exemplifieras i avsnitt 3.

Tre viktiga trender har samverkat till att området de senaste åren upplevt ett kraftigt uppsving:

- Tillgång till ökad beräkningskraft, bland annat genom utvecklingen av de kraftfulla grafikprocessorer som krävs för moderna datorspel,
- Tillgång på data har ökat: digitalisering av bilder, video, röst och text har tillgängliggjort stora datamängder som är lämpliga för maskininlärning, och
- Algoritmutvecklingen har fortsatt och tillgången till öppna arkitekturer har ökat avsevärt.

Dessa tre samverkande trender började kring 2012 att ge effekter som gjort att förmågor och tillämpningar inom AI och maskininlärning accelererat kraftfullt. Framgångarna har i sin tur lett till ökad finansiering inom området.

2.1 Olika målsättningar

Den övergripande målsättningen och beskrivningen av vad man egentligen håller på med skiljer sig för olika intressenter, exempelvis utvecklare, forskare, produktägare, försäljare, och mellan användare av olika AI-ansatser. För att förstå en AI-ansats eller ett aktuellt projekt behöver man därför förstå olika AI-ansatsers grundläggande målsättningar.

Ett klassiskt sätt att dela upp den övergripande målsättningen är att skilja mellan stark och svag AI. Med stark AI avses en målsättning där man vill uppnå mänslig nivå av intelligens, medan man med svag AI avser smart och effektiv hantering av något beräkningsproblem. Ett liknande sätt att beskriva skillnaderna i övergripande mål är att klargöra om man pratar om:

- **Artificiell superintelligens (ASI)**, som avser ett superintellekt som är intelligentare än de bästa mänskliga experterna inom ett eller flera områden.
- **Artificiell generell intelligens (AGI)**, som avser maskiner vars intelligens kan användas till alla möjliga problem och uppgifter, motsvarande de som en människa kan hantera (ibland används begreppet Human Level Intelligence).
- **Artificiell smal (narrow) intelligens (ANI)**, som avser maskinintelligens avsedd för en avgränsad, specifik uppgift.

Forskare inom fältet ser ofta sin forskning som exempel på smal/svag AI, men tolkas av allmänheten utanför som försök till superintelligens. Dessutom ska man komma ihåg att även om forskaren i praktiken ser sig som utvecklare av artificiell svag/smal intelligens, så är de också intresserade av artificiell generell intelligens.

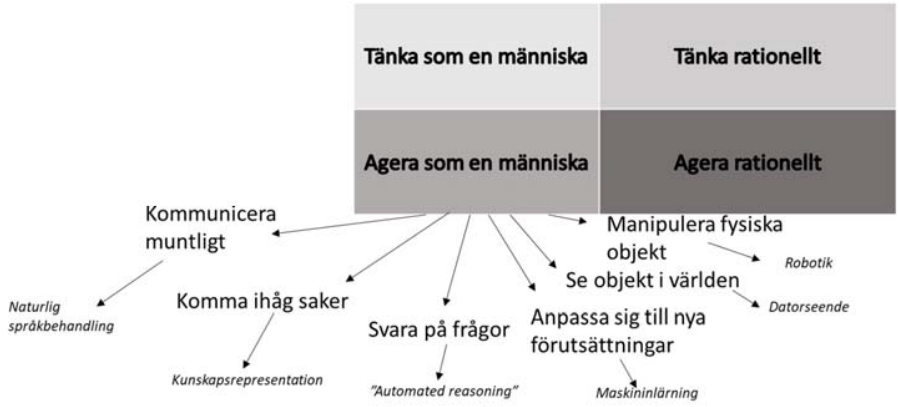
Olika typer av problem är olika svåra att lösa för AI-baserade system. Att känna igen och greppa olikformade klossar med en gripklo, med förmåga motsvarande den hos ett några år gammalt barn, har till exempel varit ett svårare beräkningsmässigt problem än att slå människan i schack, som exempelvis när Deep Blue datorn slog Kasparov i schack 1996. Svårigheten beror i hög utsträckning på hur väldefinierade reglerna är, storleken/förgreningsfaktorn på beslutsrymden och mängden informationsbrus.

Russel och Norvigs (2010) försök att beskriva olika målbilder för utvecklingen återfinns i Figur 1 och används här som en central beskrivning av ett sätt att dela upp ambitionen för en AI-ansats eller projekt:

Figur 1. Russel & Norvigs (2010) sätt att skilja mellan AI-ansater m.a.p. deras ambitionsnivå.

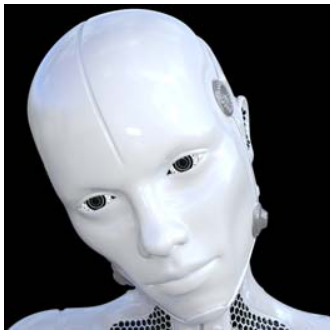
Russel och Norvig skiljer på en ledd mellan att tänka som en människa, där AI-ansatsen hämtar inspiration eller anstränger sig för att ha en beslutsprocess som liknar en människas, jämfört med om tänkandet är rationellt. Här är det rationella resultatet avgörande, inte om en människa skulle lösa problemet på samma sätt. På den andra ledden handlar det som att agera i världen som en människa, alternativt agera på ett rationellt och effektivt sätt.

I Figur 2 så exemplifieras ytterligare förmågor som ett system måste ha om det ska kunna agera i världen på samma sätt som en människa. För att kunna agera i världen blir alltså delområden som kunskapsrepresentation, inlärningsförmåga, mönsterigenkänning, sökfunktioner/optimering, resonerande, datorseende, behandling av naturligt språk, reglerteknik, och robotik centrala (notera att flera av dessa förmågor behövs även i de andra delarna av fyrfältaren). Inom alla de områden som är kursiverade finns utbredd AI-forskning.



Figur 2. Russel och Norvigs (2010) uppdelning, med fördjupning av förmågor som behövs för att kunna agera som en människa.

Kopplingen till robotiken och fysiska system är något som också kan påverka uppfattning om vad AI är. På första sidan i många presentationer rörande AI så förekommer ett relativt människoliknande futuristiskt robothuvud som i Figur 3, troligen eftersom det skapar en intresseväckande bild. Det kan dock snedvridna uppfattningen eftersom ett system mycket väl kan vara ett exempel på användning av AI, utan att nödvändigtvis vara manifesterat som en robot med humanoida drag.



Figur 3. Exempel på robothuvud.

Ett aktuellt exempel är roboten Sophia[3] i Figur 4, tillverkad av företaget Hansom Robotics. Roboten aktiverades 2016 och har visats på flera mässor och TV shower. Personer som granskat Sophia öppna källkod karakteriserar roboten som en relativt enkel så kallad chatbot, med ett robotansikte, som använder en rad olika AI-tekniker. Som ett PR-trick fick Sophia saudiskt medborgarskap under 2017 och Sophia nämns därför här som ett tydligt exempel på hur det snabbt kan bli svårt med definitioner och gränsdragningar. Eftersom Sophia har medborgarskap bör hon rimligen betraktas som en intelligent varelse med en medborgares rättigheter, samtidigt som Sophias förmåga till agerande i många avseenden ligger långt från vad en människa kan hantera och prestera.

Figur 4. Roboten Sophia.

Synen på och förståelsen för vad AI är påverkas också lätt av populärkultur och Science Fiction litteratur/filmer där robotarna i Star Wars-filmerna, robotarna i Terminator-filmerna, eller HAL i 2001-filmen alla är exempel. AI är i populärkulturen ett tacksamt och intresseväckande område, men visionerna ligger ganska långt från vad som är möjligt idag. Tillämpningar som Deepfake[4], Face2Face[5], Lyrebird[6], och DeepAngel[7], som alla på olika sätt kan användas för att skapa falsk information är exempel på aktuella tillämpningar som påverkar uppfattningen om vad AI är kapabelt till. Ett annat exempel är den i Kina populära chatboten Xiaoice[8], med 600 miljoner registrerade användare, som har en uppfattning om människors känsloläge och som kan generera dikter. Den kinesiska nyhetsbyrån

3 https://en.wikipedia.org/wiki/Sophia_(robot)
4 https://en.wikipedia.org/wiki/Deepfake – AI-tekniker används för att skapa filmer med falska ansikten
5 https://www.youtube.com/watch?v=ohmajJTcpNk – AI-tekniker används för att styra ansikten
6 https://www.youtube.com/watch?v=YfU_sWHT8mo – AI-tekniker används för att skapa kopior av röster
7 http://deepangel.media.mit.edu – AI-tekniker används för att ta bort föremål ur bilder
8 https://en.wikipedia.org/wiki/Xiaoice

Xinhua lanserade nyligen också virtuella nyhetsuppläsare[9] med utseende, läppsynkning, och rörelsemönster modellerat från riktiga nyhetspresentatörer.

AI-området inbjuder till en mängd i princip filosofiska funderingar om vad intelligens är och om människor ska betraktas som intelligenta. Intelligens i sig är ett begrepp som är svårt att definiera, men innehåll som ofta återkommer är självmedvetande, förståelse, att kunna uppfatta och härleda information, att lära sig fort, att lära sig av sina erfarenheter, resonerande, problemlösning, planering, kreativitet, förstå komplexa idéer och samband, tänka abstrakt, förmåga att adaptivt hantera en föränderlig omvärd och föränderliga arbetsuppgifter samt att anpassa sin kunskap för att möta sina mål. Intelligens inbegriper en bred och djup förmåga att "förstå vad som är på gång", förstå sammanhang, och räkna ut hur man ska agera, även med ofullständig information.

Ett klassiskt tankeexempel rörande intelligens är Searles kinesiska rum (Searle, 1980) där, förenklat beskrivet, en person sitter i ett rum och genom en lucka får in texter på kinesiska och översätter dem steg för steg med hjälp av en instruktionsbok. Personen i rummet förstår inte innebörden i vad som översätts utan bara följer instruktionerna i instruktionsboken. Med hjälp av instruktionerna lyckas människan dock producera en översättning som av en kines uppfattas som att vara på kinesiska. Kan personen då kinesiska och kan betraktas som intelligent? Eller är det rummet (med människan och instruktionsboken) som system som är intelligent? Eller är det personen som skrev instruktionsboken som är intelligent? Finns det någon skillnad på "sann intelligens" och vad som utifrån kan upplevas som intelligent beteende? Med dagens automatiserade översättingstjänster som exempelvis Google Translate är Searles tankeexperiment inte särskilt långsökt. Tidigare ansåg man att förståelse för språkets innehåll krävdes för att göra maskinöversättning, men för många praktiska tillämpningar visar det sig vara fullt tillräckligt att processa några miljoner meningar för att statistiskt se i vilka sammanhang ord förekommer och sedan göra översättningen utifrån detta, utan förståelse för det språkliga innehållet.

Ett annat ofta använt begrepp är det så kallade Turingtestet, som har föreslagits som ett test på ett AI-systems intelligens. I ett Turingtest får en mänsklig bedömare skicka skrivna meddelandet i en chat-miljö till två svarsgivare, en mänsklig och en baserad på AI. Om bedömaren utifrån sina frågor och svarsgivarnas svar inte kan avgöra vilken av svarsgivarna som är vem, sägs AI-systemet kunna uppvisa mänsklig intelligens. Turingtestet är ett relativt enkelt förslag på test och om Turingtestet är kriteriet så måste nog AI anses redan ha nått mänsklig intelligens.

Utan att i denna rapport slutgiltigt definiera intelligens så är det intressant att peka ut att det många AI-forskare egentligen eftersträvar är rationalitet, med vilket avses förmågan av att välja en "bästa handling" givet ett specifikt mål, några optimeringskriterier, och tillgängliga resurser. De har alltså inte nödvändigtvis ambitionen att utveckla system och algoritmer som täcker alla aspekter av mänsklig intelligens.

AI-baserade lösningar går att tillämpa på många verksamhetsområden, så fort det finns beräkningsbara problem och en tillräcklig mängd data att analysera. Ett antal exempel listas i avsnitt 3. AI blir dock ofta svårt när problemet eller målet är otydligt och om analysen använder en för liten eller oorganiserad datamängd. En relativt stor del av tiden i många AI-projekt går därför ut på att samla in och strukturera upp data. AI-lösningar kan dock vara onödiga om det går att hitta tillräckliga svar med andra metoder, exempelvis om något beräkningsproblem redan karakteriseras av att det finns tydliga regler och kända gränsvärden som enkelt kan definieras.

9 https://www.bbc.com/news/technology-46136504

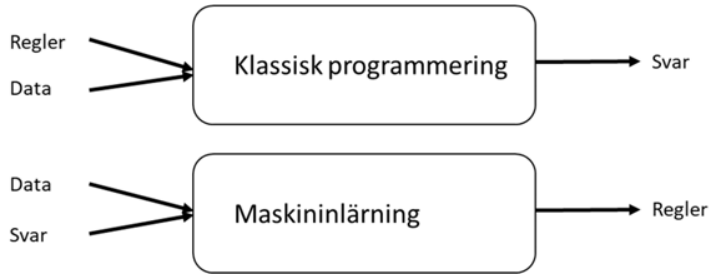
Kärnan i AI är användandet av olika mer eller mindre komplexa och ofta statistiskt baserade algoritmer, som hanterar en datamängd som är läsbar av en dator. Denna datamängd kan komma från många olika typer av digitala källor som databaser, text, kartor, bilder, tal och strömmande video eller annan sensordata. Det är därför väl värt att påminna sig om att algoritmerna inte har någon uppfattning om datas kontext eller innebörd, om den inte anges specifikt genom mer data. Det finns alltså ingen direkt förståelse för datas innebörd, utan det handlar om många steg av hantering av data genom diverse regler och algoritmer. AI-baserade system har därför inte av sig själv förmåga att kunna göra rimlighetsbedömningar avseende sina slutsatser eller hanteringen av data.

2.2 Vilken typ av AI är det?

Närhelst man behöver analysera ett AI-projekt eller en produkt som påstås vara driven av AI så måste man alltså ställa sig nyckelfrågan: vilken typ av AI avses?

2.2.1 Maskininlärning

Ett viktigt begrepp i sammanhanget är om systemet har förmåga till maskininlärning *(machine learning)*. Med maskininlärning avses att algoritmerna som används inom aktuell AI-ansats kan utvecklas över tiden, utan att utvecklaren explicit behöver beskriva hur de ska förändras. Begreppet maskininlärning användes redan 1959 av Artur Samuel som var en tidig pionjär. Han beskrev det som ett forskningsområde som ger datorer förmågan att lära sig utan att explicit programmeras. Maskininlärning representerar således ett annat paradigm än den klassiska programmeringen, se exempelvis Chollets (2017) beskrivning i Figur 5. I klassisk programmering så skriver utvecklaren kod som beskriver de regler som behövs för att hantera indata för att få önskat svar som utdata. I maskininlärning (åtminstone inom övervakad inlärning, se avsnitt 2.5.1) så får systemet data och svar, och systemet identifierar utifrån detta de regler som behövs. Dessa regler kan sedan appliceras på nya indata för att hantera en ny datamängd och ge korrekta svar. En maskininlärningstillämpning programmeras alltså inte explicit avseende sitt beteende, utan tränas genom exponering av träningsdata.



Figur 5. Skillnaden mellan klassisk programmering och maskininlärning enligt Chollet (2017).

Nuförtiden avses ofta användningen av någon form av artificiella neurala nätverk som tränas på stora datamängder när man nämner maskininlärning, där nätverkets vikter gradvis anpassas under träningen för att optimera systemets prestation, exempelvis för en klassificeringsuppgift.

2.2.2 Symboliska och subsymboliska ansatser

En annan större skiljelinje som kan användas för att beskriva AI-området är att skilja mellan två typer av paradigm som kan kallas symboliska respektive subsymboliska ansatser.

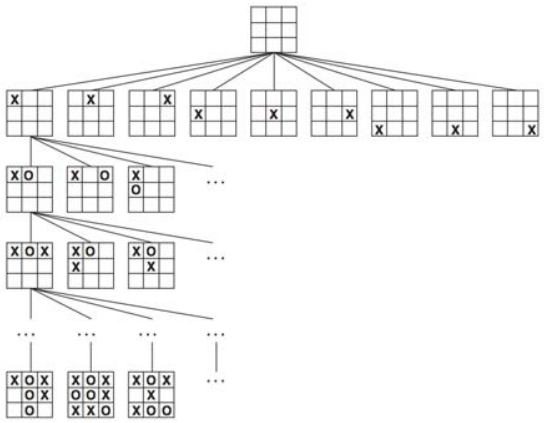
De kanske mest klassiska AI-baserade systemen, s.k. expertsystem, är regelbaserade system som använder sig av symbolisk representation. Utvecklarna skriver kod för lösning av problem på ett sätt som är relativt enkelt tolkningsbart för människor. Förenklat uttryckt så består koden av bokstäver och siffror (d.v.s. symboler) som bygger upp variabler och regler för informationshanteringen. Ett

enkelt exempel skulle kunna vara reglerna i punktlistan nedan som beskriver ett expertsystems logik för ett system som kan spela tre i rad.

1. Om din motståndare har två i rad, spela ut på den återstående rutan i raden.
2. Annars, om det finns en ruta som skapar två rader av två i rad, välj detta.
3. Annars, om mittenrutan är ledig, spela ut där.
4. Annars, om motståndaren har spelat i en hörna, spela i motstående hörna.
5. Annars, om det finns en tom hörna, spela där.
6. Annars, spela på någon tom ruta.

Begränsningen hos expertsystem är att all kunskap och regler för hur information ska hanteras måste specificeras och kodas in i systemet av utvecklarna, vilket relativt snabbt kan bli ohanterligt. För ett enkelt spel som tre i rad är det ingen svårighet att beskriva de optimala handlingarna. Schack går också att hantera, men så fort det introduceras större osäkerheter, varierande förutsättningar, och hög komplexitet så blir det praktiskt omöjligt att lösa problemen med de mer klassiska regelhanterande systemen. Fördelen med denna typ av symboliska regelhanterande system dock är att det går att följa hur systemet fattar beslut och systemet behöver inte tränas innan det kan användas.

Mycket av den tidiga AI-forskningen var fokuserade på att söka efter den bästa vägen genom en stor tillståndsrymd. Ett enkelt exempel visas i Figur 6, återigen med ett enkelt tre i rad exempel, där utsnitt ur sökrymden visas, och där målet är att söka igenom den tänkbara tillståndsrymden för att hitta ”den bästa” lösningen.



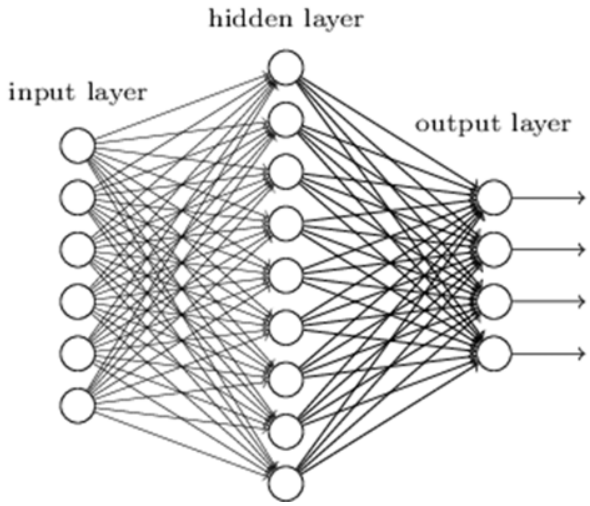
Figur 6. Sökning genom en tillståndsrymd för ett ”tre i rad” spel.

En av orsakerna till att AI omnämns så pass mycket som det gör idag, både i dagspress och i vetenskaplig litteratur, är de stora genombrott som gjorts det senaste decenniet inom det delområde som ofta kallas djupinlärning *(deep learning)*, se vidare i avsnitt 2.6. Djupinlärning bygger vidare från idéer för artificiella neurala nätverk (*artificial neural networks, ANN)*, som använder sig av subsymbolisk representation av kunskap och ibland används namnet konnektionism. ANN är något det har teoretiserats och utvecklats kring länge, men de senaste årens algoritmutveckling har tillfört flera nya förmågor som nu gör dem ännu mer praktiskt användbara. Medan det för de symboliska ansatserna finns en tydlig betydelse i varje symbol som tolkas av systemet, blir denna koppling till mening mycket mer abstrakt i de subsymboliska ANN-ansatserna.

Notera att stora delar av fortsatt fördjupning i rapporten kommer beröra de subsymboliska konnektionistiska ansatser, då denna del av AI-området haft störst genomslag de senaste åren.

Omfattande verksamhet, både avseende teori och praktisk tillämpning, finns även inom övriga AI-områden som inte behandlas i rapporten.

Ett ANN består i sin enklaste form av tre lager, ett indata-lager som tar emot data, ett dolt lager och ett utdata-lager som ger resultaten, se Figur 7. Varje nod i indatalagret söker efter ett visst särdrag (feature), se stycke 2.3, exempelvis så kanske en av noderna aktiveras av horisontella streck. Om denna nod registrerar ett horisontellt streck i datamängden så aktiveras noden och skickar sin aktivering vidare till alla noder den är kopplad mot.

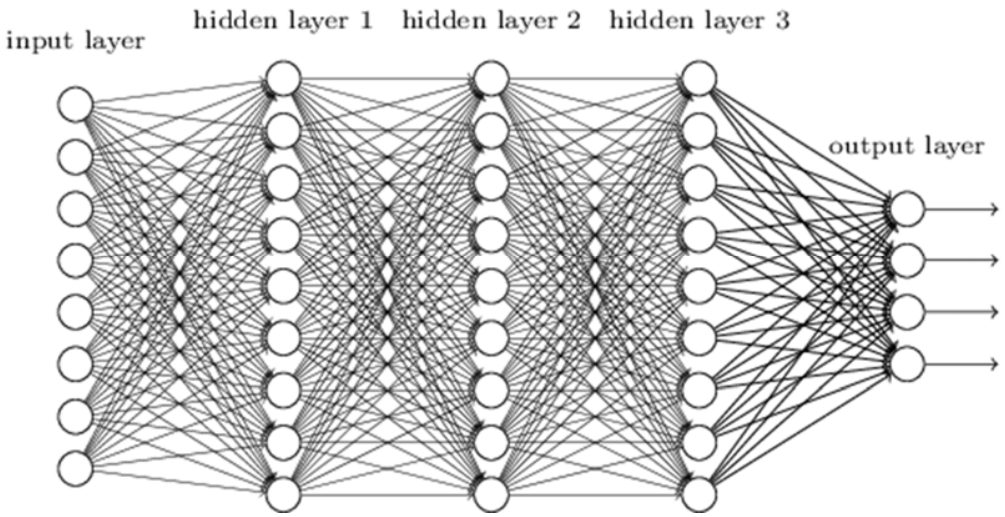


Figur 7. Enkel schematisk beskrivning av klassiskt neuralt nätverk (feed-forward).

Några grundläggande parametrar för ett ANN är:

- Antal noder i inlagret
- Antal noder i utlagret
- Antal dolda lager och antalet noder i dem
- Aktiveringsvikter och aktiveringströskeln mellan länkade noder i nätverket
- Topologin/nätverksstrukturen, d.v.s. vilka noder som länkar till vilka
- Om noder kan återmata information till sig själv, s.k. *recurrent networks*
- Om information rörande felklassificering återmatas till nätverket, s.k *backpropagation*

Allt eftersom beräkningskraften ökat så har antalet dolda lager och antalet noder per lager ökat. Ett något mer omfattande nätverk, med fler dolda lager, syns i Figur 8. Dagens neurala nätverk, särskilt de som används inom djupinlärning, kan ha hundratals dolda lager.



Figur 8. Schematisk beskrivning av ett klassiskt feed-forward neuralt nätverk med fler lager. Observera att alla noder enbart kopplar framåt (d.v.s. åt höger i figuren) och kopplar till alla noder i nästa lager.

Komplexiteten, de relativt abstrakta koncepten, och terminologin rörande ANN gör det snabbt oöverskådligt för den oinvigde. Förståelse för hur lager, vikter, uppdatering av vikter och motsvarande fungerar förklaras mycket pedagogiskt och animerat på 3Blue1Brown[10] vilken rekommenderas för läsaren för att snabbt förstå grunderna. Filmen förklarar den mest klassiska versionen av *feed-forward* nätverk, men förståelse för deras uppbyggnad och funktion utgör basen för att kunna förstå de senaste versionerna av nätverk som används i dagens djupinlärningssystem.

Processen kring maskininlärning med ANN ser typiskt ut enligt följande punkter:

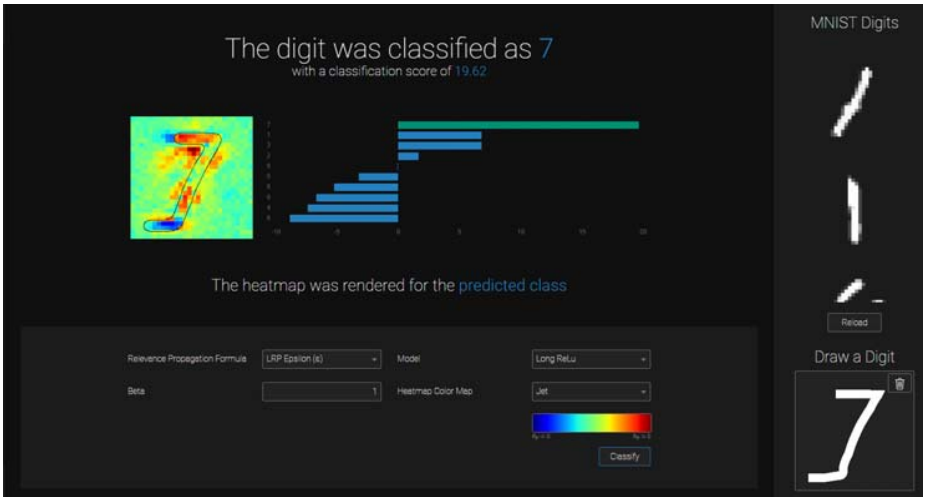
1. Samla data,
2. Städa och organisera datamängden,
3. Förstå datamängden,
4. Bygg en modell (d.v.s. en nätverksarkitektur),
5. Låt modellen tränas på data,
6. Justera modellen,
7. Validera modellen på en tidigare oanvänd delmängd av data, och
8. Använd det tränade nätverket/modellen för den tilltänka uppgiften.

Att samla in, städa, och förstå datamängden kan utan problem uppta 70% av tiden i ett projekt och denna aktivitet behöver ofta stöd av domänexperter som kan beskriva vad olika variabler representerar och hur/varför det som verkar vara avvikelser i data kan vara enkelt förklarat om man förstår domänen.

En praktisk, upplevelsebaserad förståelse av ANN kan erhållas genom att utforska några exempelapplikationer som finns tillgängliga via fotnot[11]. Dessa applikationer visar klassificering av siffror, bilder och text. Exemplet där man kan experimentera med igenkänning av siffror visas i Figur

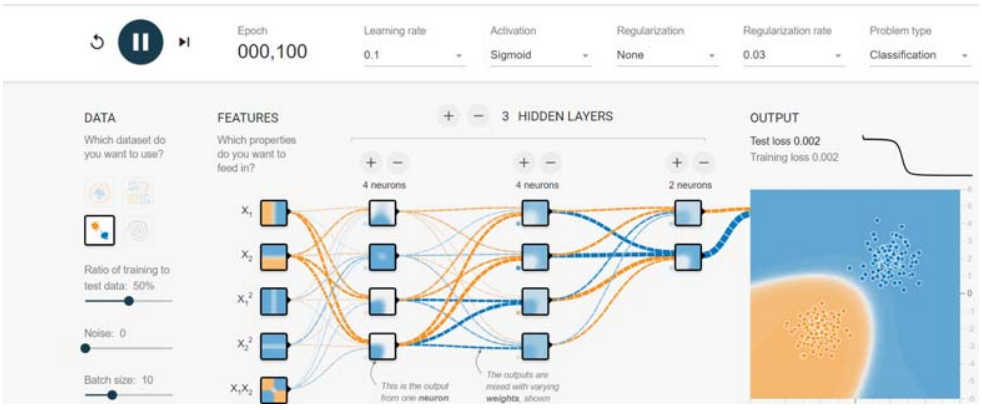
10 https://www.youtube.com/watch?v=aircAruvnKk
11 https://lrpserver.hhi.fraunhofer.de

9. Demo av sifferigenkänning. I denna demo kan användaren rita en siffra i rutan längst ner till höger och nätverkets klassificering samt säkerhet i klassificeringen presenteras.



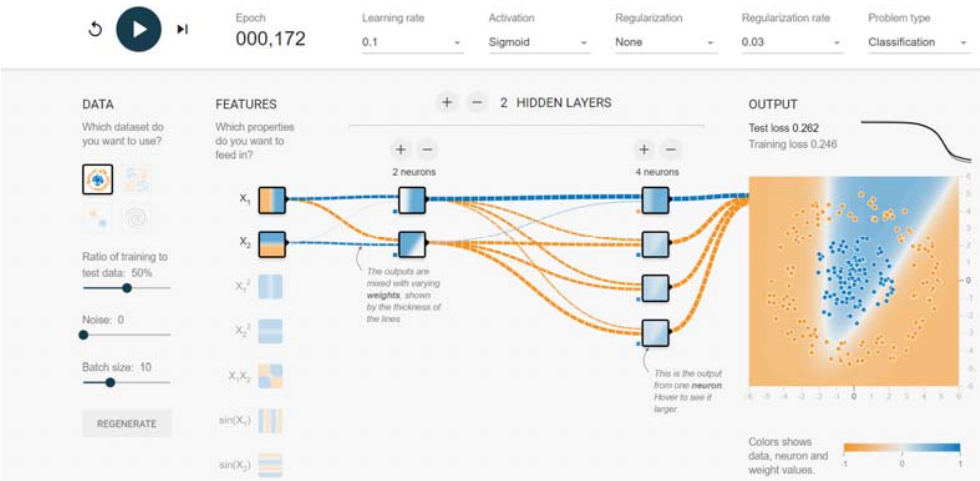
Figur 9. Demo av sifferigenkänning.

En annan, lite mer krävande, resurs för att visuellt och interaktivt uppleva hur olika egenskaper hos ett ANN påverkar hur nätverket lyckas klassificera, är ett exempel i arkitekturen TensorFlow[12]. Genom att variera de olika inställningarna i gränssnittet i denna demo ser man i *output*-rutan hur nätverket lyckats klassificera de olika prickarna, d.v.s. den yta som efter ett antal körda epoker/iterationer visar en viss färg bör innehålla alla prickar som har den färgen. Nätverket har då korrekt lyckats klassificera dem. Olika antal noder och lager som behövs för att korrekt klassificera i de olika datamängder som kan väljas som inparametrar för att sedan få uppleva hur nätverkets förmåga varierar med det. Exemplen som anges här är generella, men för en kärnkraftsintresserad läsare kan prickarna få representera variabler i olika systemtillstånd som ett AI-baserat system ska klassificera som önskvärda eller icke önskvärda.



Figur 10. Exempel på lyckad klassificering efter 100 iterationer (epochs), dvs. alla gula prickarna befinner sig inom det gulmarkerade området.

12 https://playground.tensorflow.org



Figur 11. Exempel på mindre lyckad klassificering efter 172 iterationer.

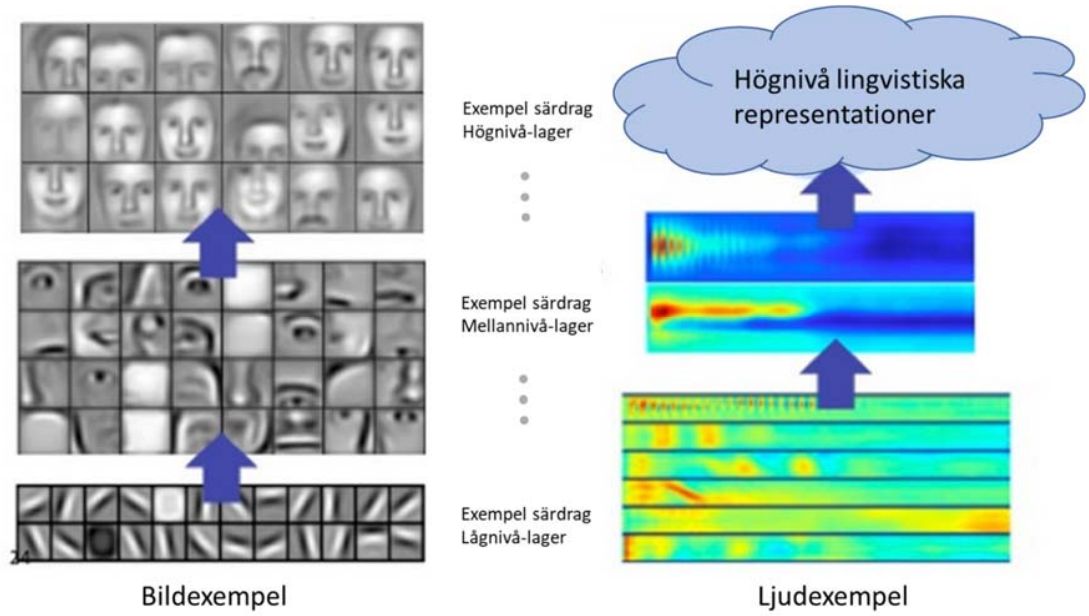
Artificiella neurala nätverk är framför allt mönsterigenkännare som kvantifierar relationerna mellan olika särdrag i den datamängd som nätverket tränas på. Förklaringen av hur systemet klassificerar den indata systemet hanterar är ofta mycket svårt att tillgodogöra sig för en människa. Ett artificiellt neuralt nätverk av idag kan innehålla miljontals parametrar som uppdateras, och efter träning kanske resulterar i en lösning som fungerar för den aktuella klassificeringsuppgiften, men som inte nödvändigtvis är den enklaste eller mest effektiva lösningen.

För att ett AI-system skall bli effektivt behövs ett gränssnitt mot andra system/omvärlden, stora mängder träningsdata, och en analyskomponent baserad på någon AI-ansats. Utan tillgång till en stor och lämplig datamängd kan inget projekt som använder sig av någon ANN-baserad ansats lyckas. Historiskt har tillgången till dessa datamängder varit en tydligt begränsande faktor och relativt mycket energi har gått åt till att skapa dessa datamängder. Data från en skarp/verklig situation innehåller ofta relativt mycket brus vilket gör att data behöver städas och struktureras. När en lämplig datamängd finns tillgänglig kan dock olika maskininlärningsansatser användas för att hitta tidigare okända samband eller mönster i datamängden, hitta avvikande datapunkter, identifiera trender, och predicera utfall. Användaren måste dock alltid komma ihåg att mönsterigenkänningen baserar sig på data som kommer från en tid och händelser som redan skett. Man bör också inse att kraften i stora datamängder är avsevärd, men att de värdefulla insikter en människa kan göra ofta kan basera sig på små datamängder, där data verkligen är gällande för de aktuella förutsättningarna.

För att göra området ännu svårare att överblicka så är det inte ovanligt att det i tillämpade produkter finns komponenter som använder sig av olika AI-ansatser, med flera symboliska respektive varianter av subsymbolisk hantering och/eller olika träningsansatser för olika delfunktioner i systemet. Detta bidrar till att det kan vara svårt att göra helt rena klassificeringar av system. Att endast använda subsymbolisk maskininlärning, oavsett hur revolutionerande och framgångsrik den är, bedöms inte vara tillräcklig som metod för många beslutsstöd. Det är helt nödvändigt att ta resultaten från subsymboliska ansatser och allt under den nivå där människor medvetet kan tänka och förstå beslutsfattandet upp till den symboliska nivån och inkludera dem i beslutsstödssystem. De sammanlagda resultaten kan då sammanställas för att presenteras och visualiseras för mänskliga beslutsfattare på ett begripligt sätt.

2.3 Särdrag

Ett centralt begrepp i alla neurala nätverk är identifikationen eller extraktionen av särdrag *(features).* Processen påbörjas vid det första indatalagret och sker sedan successivt genom de dolda lagren. Noderna i indatalagret är programmerade att aktiveras när de ”ser” sitt särdrag i indatamängden. I början av en särdragsidentifieringsprocess är det mycket enkla särdrag. För ett visuellt baserat exempel skulle särdraget kunna vara om det finns något i pixel 1:1 av en bild på 100*100 pixlar. Om det finns något där så aktiveras denna nod och skickar sin aktivering vidare. Genom behandling av flera lager bildas så mer och mer meningsbärande högnivårepresentioner som eventuellt börjar närma sig de som vi människor medvetet kan använda för att beskriva och förstå en bild. Se Figur 12 för exempel från bildanalys och ljudanalys. Observera dock att det finns flera lager både före lågnivålager-exemplet och mellan övriga efterföljande lager i exemplen nedan.



Figur 12. Exemplifiering av särdragsidentifiering. !

2.4 Nätverksstruktur

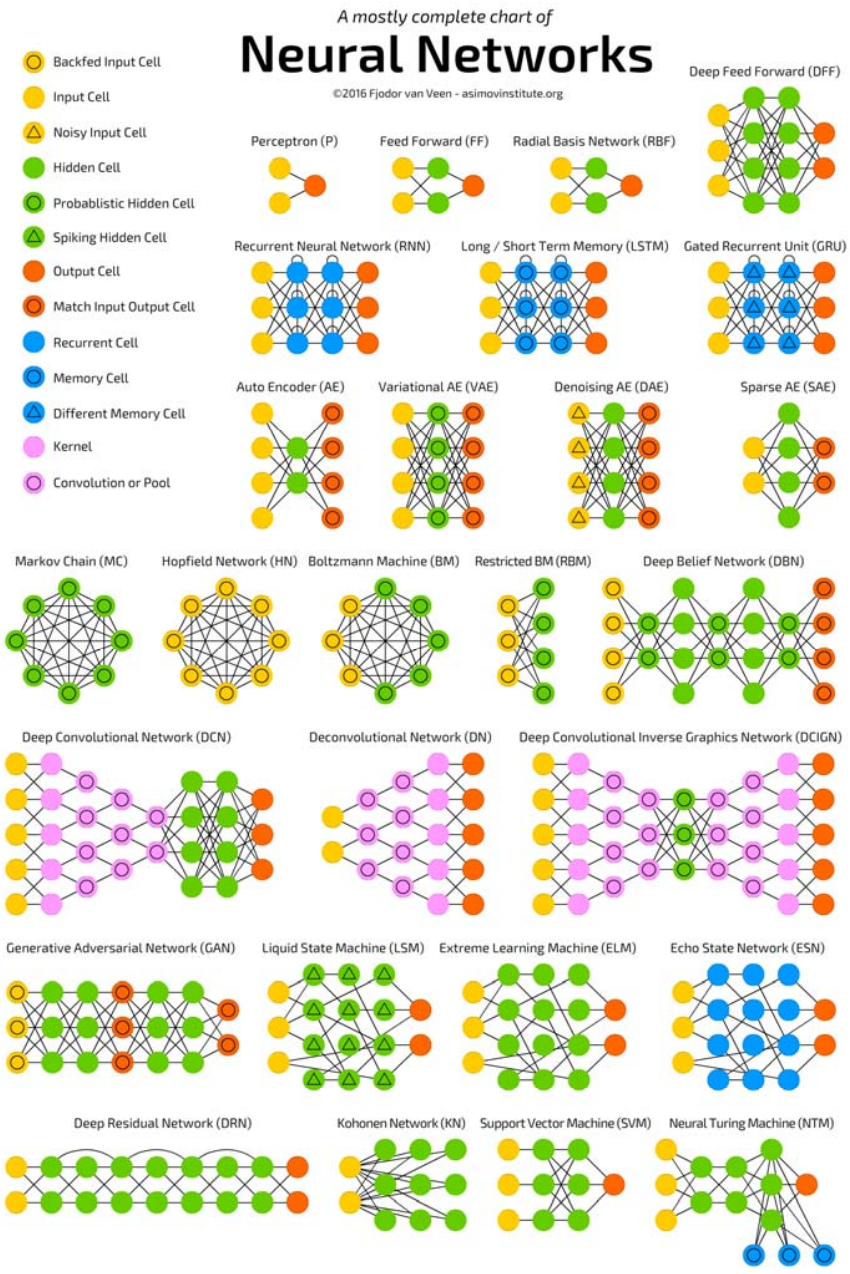
Det finns en mängd varianter på artificiella neurala nätverk och några vanliga exempel är:

- Feed-forward neural networks
- Recurrent neural networks
- Multi-layer perceptrons (MLP)
- Convolutional neural networks
- Long /Short term memory networks (LSTM)
- Recursive neural networks
- Deep belief networks
- Self-Organizing Maps
- Support Vector Machines
- Boltzmann machines
- Auto-encoders

Feed forward neural networks (FFNN) som beskrivs i avsnitt 2.2.2 är den klassiska typen av ANN där information bara skickas framåt inom det neurala nätverket. En annan vanlig typ av ANN är så kallade återkopplande *Recurrent Neural Networks* (RNN). Med RNN avses nätverk där information även tillåts skickas bakåt inom nätverket och noderna kan tillåtas att ha loopar. För att höja prestandan i en klassificeringsuppgift används idag ofta så kallad *backpropagation* och med detta avses att nätverket gör omtag och återmatning av fel när vikterna mellan noderna i nätverket uppdateras, se om önskat vidare förklaring via fotnot [13] [14]. Det finns även fler parametrar som kan varieras, exempelvis aktiveringströsklar och aktiveringsfunktioner, men det beskrivs inte vidare i denna rapport.

13 https://www.youtube.com/watch?v=Ilg3gGewQ5U
14 https://www.youtube.com/watch?v=vhAt4EoBJhc

Dessa olika nätverk skiljer sig avseende flera egenskaper, men en viktig skillnad är själva strukturen i nätverken, d.v.s. vilka noders aktivering som påverkar andra noders aktivering samt användning av noder med olika egenskaper. Van Veen (2016)[15] presenterar kartan i Figur 13 som övergripande beskriver många olika typer av ANN.



Figur 13. Karta över olika typer av neurala nätverk enligt VanVeen (2016).

15 http://www.asimovinstitute.org/neural-network-zoo

2.5 Träningsmetoder

En central skiljelinje mellan de olika subsymboliska ansatserna är hur nätverken tränas, d.v.s. hur vikterna mellan olika noder i nätverket uppdateras efter att nätverket exponerats för indata. En ofta använd uppdelning är att skilja mellan följande tre huvudtyper av träning:

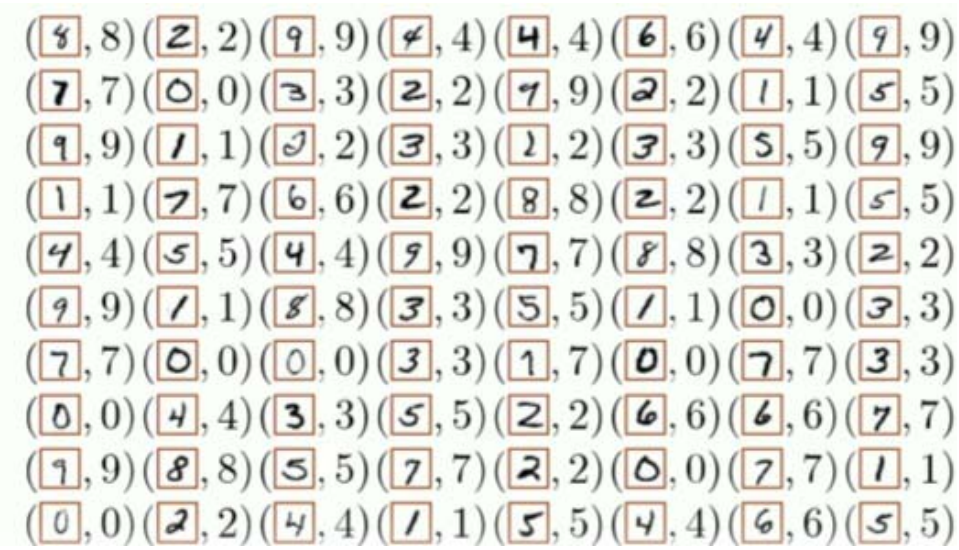
- Övervakad inlärning *(supervised learning)*
- Oövervakad inlärning *(unsupervised learning)*
- Förstärkningsinlärning *(reinforcement learning)*

2.5.1 Övervakad inlärning

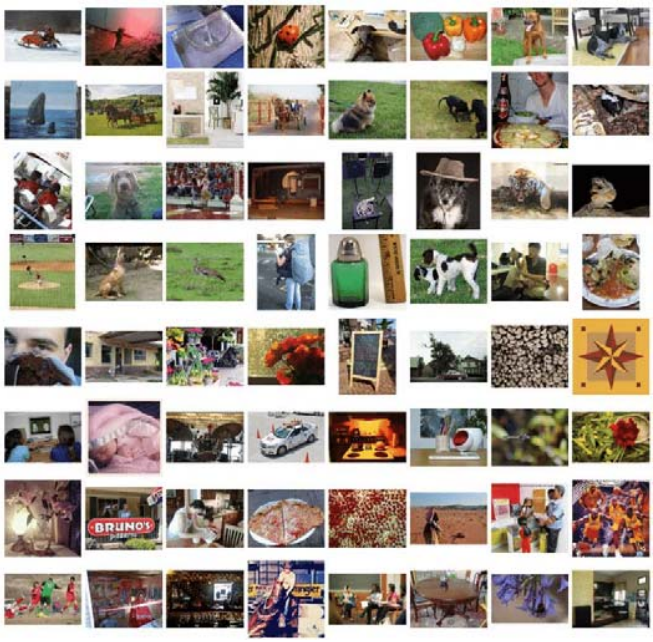
Övervakad inlärning *(supervised learning)* är den klassiskt sett mest använda versionen av träning och bedömt fortfarande den mest använda formen. Nätverket lär sig här rätt svar för varje stimuli från någon annan, oftast ett av människor annoterat rätt svar. För varje stimuli, exempelvis en bild, har alltså en människa angett vad som finns på bilden.

Ett tydligt och klassiskt exempel visas i Figur 14. I MNIST databasen ingår 70 000 handskrivna siffror med rätt tolkning angiven, exempelvis så visar den översta vänstra cellen en handskriven åtta och rätt svar är angivet till en åtta. MNIST databasen är en ofta använd referensdatamängd när olika algoritmer och nätverksutformning ska jämföras. ImageNet är ett annat exempel på en ofta använd referensdatamängd där mer än 14 miljoner bilder annoterats, se Figur 15.

Ett exempel på övervakad inlärning hos människor kan vara ett barn som ska lära sig läsa. Oftast sker det genom att en vuxen person pekar på en bokstav och uttalar stavelsen. Barnet lär sig då med hjälp av att koppla den visuella bilden av bokstaven till ett ljud. På samma sätt behöver algoritmen tränas upp, under träningen ger man den data som är kopplad till en etikett *(label)* som avspeglar korrekt eller önskat värde för just den datapunkten.



Figur 14. Utdrag ur MNIST datamängden.



Figur 15. Exempel på bilder från ImageNET datamängden, dock utan annotering.

För att övervakad inlärning ska fungera behövs en stor mängd träningsdata, där nödvändig omfattning beror på klassificeringsproblemets komplexitet, dock oftast många tusen par av stimuli tillsammans med rätt svar.

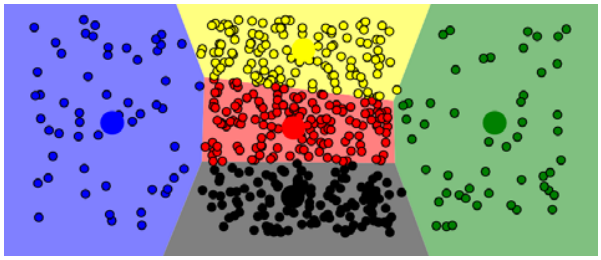
Ett ANN kommer inte på ett för människor lätt genomskådligt sätt kunna förklara varför det har klassificerat exempelvis varje bild i Figur 15 på ett visst sätt. Om sammanhanget för klassificeringen ändras, exempelvis när ett ANN utsätts för en annan typ av datamängd än vad det tränats på, eller om datamängden inte innehåller den relevanta informationen kommer nätverket att inte lyckats klassificera med någon högre precision. En viktig del av kvalitetssäkringen är därför att man först låter nätverket träna på en delmängd av tillgängliga data och sedan exponerar det för ny en ny delmängd som det tidigare inte utsatts för, för att på så sätt se hur bra den uppträdande klassificeringsförmågan går att generalisera. Värt att notera är att det finns ett antal uppmärksammade exempel på att dagens företag i relativt hög utsträckning använder sig av mänskliga kontrollanter för att finslipa och kontrollera sina AI-baserade sökalgoritmer[16].

2.5.2 Oövervakad inlärning

Med oövervakad inlärning (unsupervised learning) avses träning där det inte finns några etiketter eller korrekta svar. Det handlar istället om att hitta strukturen i datamängden och göra kluster av data som liknar varandra. Oövervakad inlärning ligger tydligt nära den klassiska statistiken med användning av statistiska metoder som *k-means clustering*, *nearest neighbour, random forest*, logistisk regression, och linjär regression. Ett interaktivt exempel på *k-means clustering* återfinns via fotnot[17] och en liknande animering finns via fotnot[18]. Om de mindre prickarna i Figur 16 exempelvis skulle vara mätvärden från någon kärnteknisk process där önskvärt intervall skulle vara kring den större röda

16 https://link.medium.com/dFZ37GI9HT
17 https://www.naftaliharris.com/blog/visualizing-k-means-clustering
18 https://www.youtube.com/watch?v=5I3Ei69I40s

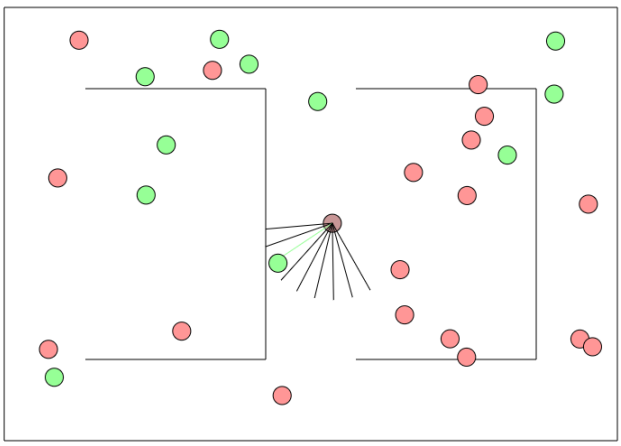
pricken skulle övriga mindre prickar som hamnar närmast någon annan av de större prickarna kunna identifieras som avvikelser och denna information kan användas för att informera operatörerna.



Figur 16. Visuellt exempel på k-means clustering.

2.5.3 Förstärkningsinlärning

En träningsmetod och typ av algoritmer som just nu får mycket uppmärksamhet är så kallad förstärkningsinlärning *(reinforcement learning)*, där AI-systemet får sekventiell återkoppling under träningen som gör att önskvärda beteenden förstärks. Förstärkningsinlärning har undersökts av forskare relativt länge, se exempelvis Lin (1993), men har nu börjat användas i högre utsträckning, ibland i kombination med djupinlärning. Förstärkningsinlärning är användbart när övervakad inlärning inte är möjligt. Fördjupning kring förstärkningsinlärning finns exempelvis tillgänglig i Sutton & Barto (2018)[19]. För att kunna ge den sekventiella återkopplingen under träningen behövs tydliga kriterier på vad som är önskvärt respektive inte önskvärt för varje handling systemet kan ta eller för slutresultatet. En mycket explicit uppfattning om målfunktionen är alltså ett krav för att förstärkningsinlärning ska vara framgångrikt. Interaktiva exempel finns här[20] [21] [22]. I exemplen finns en agent med perception (pricken i mitten med siktlinjer i Figur 17) som ska navigera i ett rum och där hitta de gröna prickarna samt undvika de röda.



Figur 17. Exempel där agent med reinforcement learning lär sig navigera och hitta ”rätt” prickar.

Förstärkningsinlärning är användbart när det finns ett flertal möjliga handlingar som tydligt kopplar till definierbara incitament eller mål. En stor fördel är att träningen sker utan behov att utvecklarna är explicita avseende vad som är önskvärt beteende, utöver mål/incitamentstyrningen. Enkelt uttryckt

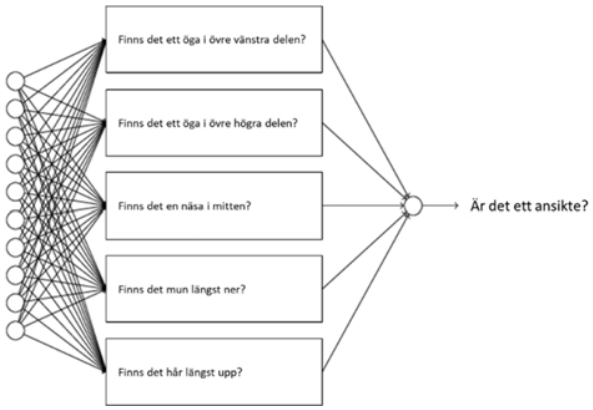
19 http://incompleteideas.net/sutton/book/RLbook2018trimmed.pdf
20 http://projects.rajivshah.com/rldemo/
21 https://www.youtube.com/watch?v=3TUZw1rlvXc
22 https://cs.stanford.edu/people/karpathy/convnetjs/demo/rldemo.html

så prövar systemet en massa handlingar, ser vad som händer och gör sedan mer av det som belönades. Förstärkningsinlärning är alltså tydligt explorativt och behöver en mycket stor datamängd att tränas på, som ofta behöver skapas genom simulering av den aktuella miljön. En nackdel är att algoritmen lär sig iterativt från början för varje problem, vilket betyder att det kan ta lång tid innan algoritmen hittar en effektiv lösning.

2.6 Djupinlärning

Djupinlärning *(deep learning)* är ett delområde inom maskininlärningen och användningen av artificiella neurala nätverk som har rönt stor uppmärksamhet de senaste åren. Fördjupningar av djupinlärningens historia och tekniska detaljer finns tillgängligt i Goodfellow, Bengio och Courville (2016)[23].

Djupinlärning bygger på hierarkisk inlärning av särdrag och ett mer förklarande namn skulle varit hierarkisk inlärning *(hierarchical learning)*. I djupinlärning försöker man inte lösa exempelvis ett klassificeringsproblem på en gång, utan tar det i flera hierarkiska steg, vilket leder till att man använder sig av flera lager som succesivt identifierar mer komplexa särdrag. Djupinlärningstillämpningarna skiljer sig från de tidigare klassiska *feed-forward* nätverken genom att de oftast använder sig av fler lager, det kan vara hundratals lager, och använder sig av mer komplexa sätt att koppla lager i nätverken till varandra. Dock är det den successiva identifieringen av särdrag som är den främsta skillnaden. Genom många lagers identifiering av särdrag blir särdragen som identifieras mer sammansatta, vilket exemplifieras i Figur 18, som skulle kunna vara ett utsnitt ur ett ANN där nätverket ska klassificera en bild. Detta lager av nätverket letar efter delar hos ett ansikte och om tillräckligt många noder aktiveras så rapporterar detta lager vidare till nästa lager att det finns ett ansikte i bilden



Figur 18. Exempel på ett lagers särdragsidentifiering som uppnås efter många tidigare lagers behandling av enklare särdrag.

System som använder djupinlärning har haft uppmärksammade genombrott de senaste tre åren, vad gäller exempelvis bild och ansiktsigenkänning, automatisk översättning, sammanfattande av rapporter, och att lära sig spelstrategier genom att spela mot sig själv. Det kanske mest uppmärksammade genombrottet var när systemet AlphaGo lyckade besegra en mycket framstående Go-mästare 2016. Go har tidigare varit för svårt att lösa på grund av den explosionsartade tillväxten av möjliga tillstånd, men här användes djupinlärning där AlphaGo efter att fått lära sig reglerna, fick observera ett stort antal partier och sedan spela mot sig själv, ca 100 miljoner partier, för att lära sig optimala strategier.

23 https://www.deeplearningbook.org

Djupinlärning kan använda sig både av övervakad, oövervakad inlärning, och förstärkningsinlärning. Deepcognition.ai[24] erbjuder en web-baserad, kostnadsfri applikation där den intresserade kan experimentera med existerande eller egna datamängder.

Möjligheten till djupinlärning bygger på vidareutvecklingar av bakomliggande algoritmer, nätverksutformning, och ökad processorkraft. I en tillämpning med exempelvis 100 lager och 100 noder i varje lager, blir det många vikter som ska räknas om när nätverket uppdateras efter varje nytt stimuli. Utvecklingen har också stimulerats av tillgången på allt större träningsmängder lättillgängliga data.

Konvolutionslager *(convolutional layers)* är en särskild sorts lager som ofta används inom djupinlärning och då kallas nätverken CNN, *convolutional neural networks*. Figur 19 visar översiktligt vad som sker i ett konvolutionslager. Ett filter (här på 3x3 celler som visas i mitten) multipliceras först med de nio rödmarkerade cellerna i matrisen märkt input, resultatet skrivs till rödmarkerade cellen i matrisen märkt Conv1. Därefter görs denna operation tills alla celler i *input*-matrisen behandlats och överförts till Conv1. Genom denna typ av transformation kan nätverken successivt utveckla större och mer meningsfulla särdrag att känna igen. Animerade exempel kan återfinnas via fotnot[25] [26].

Figur 19. Exempel på konvolution.

Yosinskys DeepVis toolbox (2013)[27] visar ett ofta refererat exempel på hur noder och konvolutionslager i en djupinlärningstillämpning aktiveras och hierarkiskt identifierar särdrag, vilket resulterar i förmåga till klassificering som sedan används för bildanalys.

Väl värt att notera att djupinlärningsfältet är tydligt lösningsorienterat och att det sällan finns tydliga teoretiska förklaringar bakom nya idéer. De flesta idéer publiceras med experimentella resultat som visar att ”det fungerar”, och det anses tillräckligt som motivering.

2.7 Andra ”heta begrepp”

Framstegen inom AI-området går mycket fort och den normala vetenskapliga publiceringsprocessen kan ofta upplevas som för långsam för att forskare ska hinna med på forskningsfronten. Det är därför inte ovanligt att AI-forskare publicerar förhandsutgåvor på exempelvis arxiv.org där det finns ett antal

24 https://deepcognition.ai
25 https://www.youtube.com/watch?v=YRhxdVk_sIs
26 https://www.youtube.com/watch?v=2-Ol7ZB0MmU
27 https://www.youtube.com/watch?v=AgkfIQ4IGaM

kategorier[28] för AI-relaterade resultat. Denna källa är dock troligen inte för rapportens tilltänkta läsare utan enbart för specialister.

2.7.1 Transfer learning

En metod som kallas *transfer learning* kan användas i djupinlärningstillämpningar där det finns för liten mängd träningsdata tillgänglig. *Transfer learning* är bedömt som nästa stora steg inom AI[29] [30]. Nätverket kan då istället tränas upp med en alternativ datamängd. Ett nätverk som tränats för en klassificeringsuppgift på en viss datamängd, där det lärt sig känna igen vissa särdrag, används sen för det egentliga klassificeringsproblemet efter viss slutträning på den nya datamängden.

2.7.2 GAN

Generative Adversial Networks (GAN) är ett begrepp som är mycket hett idag och använder sig av två nätverk som arbetar mot varandra och används i flera av de tillämpningar som demonstrerats på sistone där systemet upplevs uppvisa konstnärlig kreativitet gällande poesi eller bildgenerering, se exempelvis fotnot[31] för mer ingående beskrivning av funktionerna. Ett nätverk, generatorn, skapar utdata, exempelvis handskriven text eller en bild av ett ansikte, baserat på sin träningsdata. Det andra nätverket, diskriminatorn, utvärderar och försöker bedöma om utdata är ”verklig” baserat på samma träningsdata som det andra nätverket tränats på. Man kan jämföra med en förfalskare som genererar falska pengar vilka en diskriminerande funktion försöker avslöja, och generatorn försöker därefter förbättra sina kopior. Genom maskinlärning tränas båda nätverken att bli bättre på sin uppgift. GAN introducerades av Goodfellow m.fl. (2014) och GAN tillämpningar kan generera falska bilder och röster som är mycket svåra att avslöja. Ett exempel på hur fort fältet utvecklas framgår om man betänker vad som visades upp när GAN introducerades 2014, med vad som presenteras i Karras, Laine och Aila (2018) vad gäller skapandet av verklighetstrogna ansikten på fiktiva personer. I Goodfellow m.fl. (2014) skapades suddiga fantombilder medan det i Karras m.fl. (2018) genereras ansikten som är mycket svåra att urskilja som datorgenererade, se Figur 20.

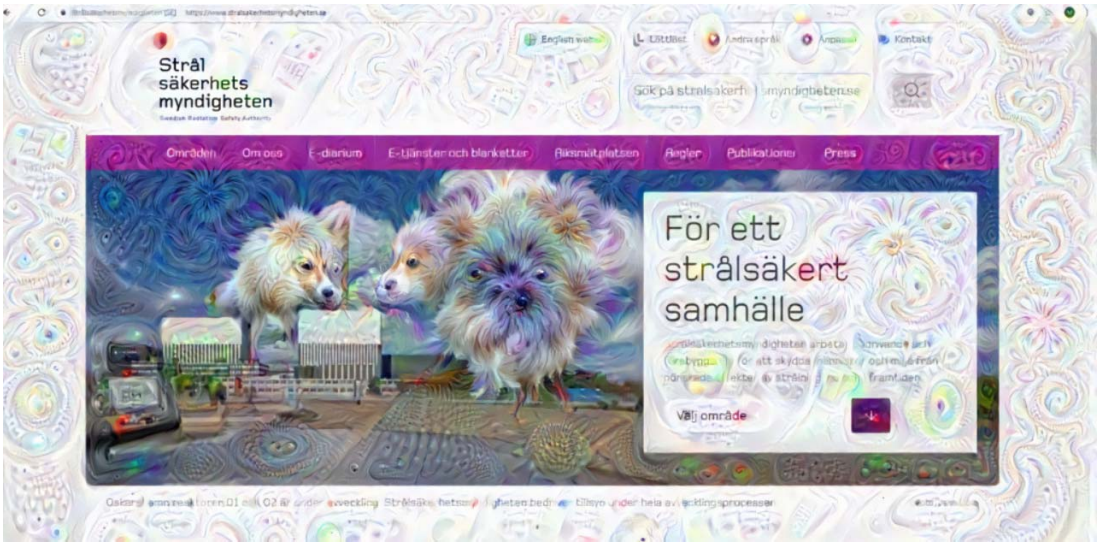
Figur 20. Jämförelse mellan GAN genererade ansikten 2014 och 2018.

ThisPersonDoesNotExist.com[32] är ett annat exempel på en GAN applikation där ett nytt datorgenerat ansikte visas upp varje gång någon visar websidan. För att mer pedagogiskt visa förmågan hos dessa

28 https://arxiv.org/list/cs.AI/recent, https://arxiv.org/list/cs.NE/recent, https://arxiv.org/list/cs.LG/recent
29 https://medium.com/owkin/transfer-learning-and-the-rise-of-collaborative-artificial-intelligence-41f9e2950657
30 https://link.medium.com/GFt3KHP9HT
31 https://medium.com/datadriveninvestor/a-leap-into-the-future-generative-adversarial-networks-96a780ed8ee6
32 https://thispersondoesnotexist.com

nätverk används i rapporten visuella bilder, men indata kan givetvis vara något helt annat, exempelvis text, ljud eller andra stimuli som kan användas för att generera dikter[33] eller filmmanus[34]. Karras m.fl. utvecklade StyleGAN[35] som det verktyg de använde för generera ansikten, men StyleGAN kan användas för att generera handskriven text, målningar och motsvarande med samma teknik. Ett nyligen uppmärksammat exempel på GAN tillämpning är GPT-2[36] från OpenAI som baserat på en stor mängd text hämtad från Internet kan generera nästa ord i en text och därigenom skapa texter som i många fall upplevs som skrivna av en människa.

Verktyg som exempelvis DeepDreamGenerator[37] kan ge en uppfattning om att AI besitter kreativitet, men när man börjar förstå vad olika lager i en djupinlärningstillämpning eller GAN gör är det inte så märkligt. Exempel på DeepDreamGenerator tolkningar av SSM websida återges i Figur 21.



Figur 21. Exempel när SSM websida modifierats av DeepDreamGenerator.

33 https://arxiv.org/pdf/1804.08473.pdf
34 https://www.youtube.com/watch?v=LY7x2Ihqjmc
35 https://www.youtube.com/watch?v=kSLJriaOumA
36 https://blog.openai.com/better-language-models
37 https://deepdreamgenerator.com/#tools

3 AI-tillämpningar

Givet att AI bör betraktas som ett forsknings- eller utvecklingsområde så kan det användas till ”vad som helst”, givet att man har ett lämpligt beräkningsproblem och en god uppfattning om vad man vill uppnå. Detta är anledningen till att AI-baserade tillämpningar går att finna i många domäner.

Chen (2016) använder fem generella kategorier, presenterade i punktlistan nedan, för att beskriva AI-ansatsers typiska problemrymd och listan ramar in många aktuella och historiska projekt på en övergripande nivå. De olika typerna av nödvändiga algoritmer som behövs i respektive problemrymd driver algoritmutvecklingen i olika riktningar.

- Logiskt resonerade
- Kunskapsrepresentation
- Planering och navigering
- Naturlig språkbehandling
- Perception

Exempel kommer ges nedan, både per typ av funktion och tillämpade exempel från branscher utanför kärnkraften, samt några exempel där AI-baserade system används i kärnkraftsspecifika tillämpningar.

Ett exempel på en tidig översikt av tillämpning av AI inom kärnkraftsdomänen ges av Uhrig (1989), se också IAEA:s konferens om AI i kärnkraften 1989[38]. Visionen och den allmänna beskrivningen av potentialen hos AI-området är fortfarande giltig, nu trettio år senare. Uhrig refererar till den klassiska typen av AI system som kallas expertsystem, men förväntningarna, visionen och föreslagna användningsområden är de samma idag med andra ansatser. Uhrigs beskrivning av potentialen med AI var följande:

> *In the operation of a nuclear power plant, great quantities of numeric, symbolic, and quantitative information are handled by the reactor operators even during routine operation. The sheer magnitude of the number of process parameters and systems interactions poses difficulties for the operators, particularly during abnormal or emergency situations. Recovery from an upset situation depends upon the facility with which the available raw data can be converted into and assimilated as meaningful knowledge. Plant personnel are sometimes affected by stress and emotion which may have varying degrees of influence on their performance. Expert systems can take some of the uncertainty and guesswork out of their decisions by rapidly providing expert advice and access to a large information base. The application of AI technologies, particularly expert systems, to control room activities in a nuclear power plant can reduce operator error and improve plant safety and reliability. Furthermore, there are a large number of nonoperating activities (e.g., testing, routine maintenance, outage planning, equipment diagnostics, fuel management, etc.) in which expert systems can increase the efficiency and effectiveness of overall plant operation.*
>
> *Other reported applications of expert systems in various stages of development include outage planning, heat rate improvement, alarm filtering, sequencing and suppression, diagnostics for instruments and equipment, welding rod selection advisor, generating welder procedure specifications that comply with regulatory codes, signal validation, disturbance analyses, condensate feedwater monitor, radwaste processing system advisor, bypass-inoperable status indicator system, sequencing BWR control rods after maneuvering, water chemistry control, pressure-temperature control during startup (to avoid pressurized thermal*

38 https://inis.iaea.org/search/search.aspx?orig_q=reportnumber:%22VTT-SYMP--109%20(VOL.%201)%22

shock problems), real-time emergency evacuation planning, and real-time radiation exposure management.

The utilities are introducing expert systems into nuclear power plants very slowly, possibly because they are reluctant to submit this new technology involving uncertainties to regulatory review until they are convinced that the benefits gained will warrant the effort required. Perhaps regulators' principal concern with expert systems is the ability to encode expertise properly, particularly the fine nuances and shades of meaning, into the knowledge base of an expert system so that it can emulate human expertise with fidelity. Another major concern is the narrow scope of the expertise and the associated limited area of applicability of expert systems. Two of the consequences of these limitations are the inability of an expert system to exhibit common sense and its limited ability to recognize when it is operating outside its field of knowledge. Researchers have sought to minimize the impact of these limitations by building "robustness" into expert systems (i.e., the ability to fail gradually and predictably when it gets outside its operating regime). These limitations, as well as the lower confidence associated with answers when data are missing or have low certainty factors, may be of concern to regulators when expert systems are introduced into the safety-related systems of nuclear power plants.

In summary, a nuclear power plant is too complex a system to be managed or operated by anyone's "gut feeling." An expert system can be the ever alert, knowledgeable assistant to the operators as well as a valuable tool for plant management. Demands for increased safety margins, lower environmental impacts, increased performance, and greater investment protection will inevitably lead to automation of most functions of nuclear power plants. In turn, automation will be paced by the ability to develop efficiently the needed software through the use of modern computer science brought about by AI programming techniques. The regulators and the public must be assured that these plants are properly designed, properly built, properly operated, and properly maintained. Artificial intelligence and expert systems can and must play a major role in providing this assurance. (Uhrig, 1989)

Schubert (2017) beskriver den aktuella utvecklingen av AI som del av beslutsstödssystem för militära ändamål. Han skiljer i sin analys mellan tre huvudtyper av beslutsstöd:

- Deskriptiva beslutsstöd – beskriver nuvarande läge
- Prediktiva beslutsstöd – beskriver vad som kommer att hända
- Preskriptiva beslutsstöd – beskriver vilka handlingsalternativ användaren har och hur man bör agera

En slutsats från Schuberts översikt är att AI idag kommer starkt på alla fronter och för alla dessa tre huvudtyper av beslutsstöd. Han beskriver exempelvis tänkt användning för prediktiva beslutsstöd baserade på AI. Utgående från en lägesbild behöver beslutsfattaren bilda sig en uppfattning om hur läget kommer att förändras i en nära framtid, samt vilka konsekvenser som kommer att uppstå på grund av förändringarna. Det inkluderar för en militär tillämpning vad en motståndare kan antas vilja genomföra och vilka möjligheter motståndaren har. För SSM kan denna typ av beslutsstöd och analyser vara användbara vid förberedelser mot antagonistiska händelser och förhållanden. Allt som är relevant för en bedömning av det framtida läget, inklusive alla de konsekvenser som följer av det nya läget ingår i analysen, på motsvarande sätt som en schackdator analyserar ”alla tänkbara” utfall. En sådan prediktiv analys görs lämpligen med ett system som samtidigt analyserar flera möjliga händelseutvecklingar. Då kan man analysera motståndarens olika möjligheter och ge varje möjlig händelseutveckling en statistisk bedömning. Eftersom det är en dator som hanterar alla de tänkbara händelseutvecklingarna behöver man inte bara begränsa sig till de mest farliga, mest gynnsamma eller

mest sannolika händelseutvecklingarna. Datorn kan internt hålla reda på hundratals olika möjligheter för att välja ut ett mindre antal hypoteser att presentera i beslutsstödssystemet. Allt eftersom situationen utvecklas kan man genomföra två olika typer av uppdateringar. För det första, så bör man kontinuerligt uppdatera sannolikheterna för över alternativa händelseutvecklingar när ny information anländer, samtidigt som man också uppdaterar den egna lägesbilden. Nya hypoteser kan tillkomma och vissa hypoteser kan tas bort. För det andra, så kan man förfina de kvarvarande hypoteserna baserat på ny inkommande information, både göra dem mer detaljerade när bättre information finns tillgänglig och dels utsträcka dem framåt i tiden. I Darling m.fl. (2018) ges exempel på AI-baserat stöd för *counterfactual reasoning* där principen för dessa system skulle kunna bli aktuellt för att hjälpa ett skiftlag i ett kärnkraftverk att hålla sig till konservativt beslutsfattande (jämför med WANO:s operatörsprinciper[39]) och inte dra förhastade slutsatser.

3.1 Tillämpning per funktion

Som antytts tidigare så kan olika AI-ansatser i princip vara applicerbara på ”vad som helst” givet att det innehåller ett passade beräkningsproblem. För problem som klustring, prediktion, regression, klassificering, rankning, summering, anomali-detektion, rekommendationer, och beslutsfattande finns många olika algoritmer från AI-området som är användbara. Goodfellow m.fl. (2016) beskriver exempelvis vanliga uppgifter för ett AI-system på följande funktionsledd:

- Classification: In this type of task, the computer program is asked to specify which of k categories some input belongs to.
- Classification with missing inputs: When some of the inputs may be missing, rather than providing a single classification function, the learning algorithm must learn a set of functions.
- Regression: In this type of task, the computer program is asked to predict a numerical value given some input.
- Transcription: In this type of task, the machine learning system is asked to observe a relatively unstructured representation of some kind of data and transcribe the information into discrete textual form.
- Machine translation: In a machine translation task, the input already consists of a sequence of symbols in some language, and the computer program must convert this into a sequence of symbols in another language.
- Structured output: Structured output tasks involve any task where the output is a vector (or other data structure containing multiple values) with important relationships between the different elements.
- Anomaly detection: In this type of task, the computer program sifts through a set of events or objects and flags some of them as being unusual or atypical.
- Synthesis and sampling: In this type of task, the machine learning algorithm is asked to generate new examples that are similar to those in the training data.
- Imputation of missing values: In this type of task the algorithm must provide a prediction of values for missing data.
- Denoising: In this type of task, the machine learning algorithm is given as input a corrupted example obtained by an unknown corruption process from a clean example. The algorithm must then predict the clean example from its corrupted version, or more generally predict the conditional probability distribution.

39 https://www.wano.info/getmedia/39dd170d-336c-4064-9cb1-51db01d3a82e/WANO-Review-2018-A4-Online.pdf.aspx

Elsevier (2018)[40] presenterar en analys där de använt AI-tekniker för att beskriva AI-området. De har analyserat 600 000 AI-relaterade dokument, från undervisnings-, forsknings-, patent- och mediakällor. Utifrån dessa dokument har de extraherat ca 800 begrepp som sedan klustrats och genom detta identifierat sju huvudsakliga kluster som beskriver AI-tillämpningsområden. De identifierade klustren är:

- Sök och optimering
- *Fuzzy Systems*
- Planering och beslutsfattande
- Naturlig språkbehandling och kunskapsrepresentation
- Datorseende
- Neurala nätverk
- Maskininlärning och sannolikhetsbaserat resonerande

Exempel på anledningar/situationer där AI är användbart är:

- Mänsklig expertis är frånvarande, exempelvis under en lång resa i rymden
- Smutsiga, repetitiva, tråkiga och farliga arbetsuppgifter
- Människor har svårt att förklara sin expertis, exempelvis språkigenkänning/förståelse eller visuell igenkänning
- Lösningen måste anpassas till varje specifikt fall/individ, exempelvis biometriska system
- Problemrymden är för stor för människans förmåga, exempelvis att läsa av alla sensorer i en stor anläggning eller analysera/kunna referera till all text som finns i böckerna på ett stort bibliotek/på Internet

3.2 Exempel på tillämpningar

Det finns en mängd exempel på tillämpade användningar av AI inom många branscher. Ett antal exempel ges nedan i en icke-uttömmande lista:

- Processautomation
- Processövervakning, avvikelsedetektion och feldiagnos
- Bild och videoanalys
- Ansiktsigenkänning
- Övervakning
- Taligenkänning/Naturlig språkbehandling
- Talsyntes
- Dialogsystem
- Textanalys/textsammanfattning
- Översättning av tal och text
- Kundtjänst
- Köpbeteende och kundanalyser
- Hälsa och sjukvårdsanalyser och prediktioner
- Biomedicinsk informatik
- Digitala assistenter
- E-post filter/IP-trafikhantering
- Robotik
- Svärmbeteenden hos autonoma plattformar

40 https://www.elsevier.com/__data/assets/pdf_file/0010/823654/ACAD-RL-AS-RE-ai-report-WEB.pdf

- *Internet of Things* styrning och analys
- Rekrytering och urval
- Finansteknologi (exempelvis trendprediktion, aktiehandel, kundutvärdering, bedrägeridetektion)
- Utbildning och undervisning
- Logistikanalys och planering
- Reseplanering
- Juridikstöd (exempelvis precedensfallsanalys)
- Fråga/svar system
- Försäkringsanalys
- Styrning av fordon (stödsystem för mänskliga förare, självkörande fordon)
- Maskinperception generellt (t.ex. datorseende)
- Grammatikkontroll
- Igenkänning av handskriven text
- Sökmotoroptimering
- Rekommendationssystem
- Underhållsplanering
- Processoptimering

Ett antal exempel på tillämpningar satta i ett kärnkraftssammanhang kan vara:

- Robotar för avfallshantering och krishantering, exempelvis Avexis[41] [42] som är en robot för uppstädning efter haveri
- Detektion av sprickor[43] [44] [45]
- Prediktion för spridning av utsläpp[46]
- Behovsdrivet effektuttag, exempelvis genom AI-system kopplade till batterilager som tar hänsyn till människors beteendemönster[47], varierande elpriser och väderprognoser
- Sammanfattningar, situationsspecifik hänvisning till rätt/relevanta stycken i styrande dokument inkl. fråga/svar tillämpningar
- Larmfiltrering/undertryckning av lägre prioriterade larm vid situationer med hög belastning för operatörerna
- Övervakning av anläggningsstatus utifrån många parametrar
- Diagnosticera avvikande förutsättningar, tillstånd, trender och transienter
- Orsaksanalys, *root cause analysis*, vid olyckor eller avvikelser
- Identifiera övergångar mellan olika driftlägen (normal, avvikelse, kritisk avvikelse)
- Analys och rekommendation av handlingar och tid/ordning för handlingar
- Underhållsanalys
- Träningstillämpningar

41 https://www.manchester.ac.uk/discover/magazine/features/robots-doing-the-dirty-work
42 https://www.researchgate.net/publication/322362014_A_Remote-operated_System_to_Map_Radiation_Dose_in_the_Fukushima_Daiichi_Primary_Containment_Vessel
43 https://www.futurity.org/deep-learning-nuclear-reactor-safety-1598922
44 https://ieeexplore.ieee.org/document/8074762
45 https://youtu.be/8O8FFey4GJo
46 https://www.nature.com/articles/s41598-018-27955-4
47 https://www.sust.se/2018/nasta-steg-for-batterilager-ar-artificiell-intelligens-ai

På en mer övergripande nivå beskriver Vinnova (2018) hur den värdeskapande AI-potentialen ligger i att:

- Automatisera funktioner i etablerade värdekedjor, verksamheter och funktioner
- Utveckla nya affärsmodeller, varor, tjänster och systemlösningar
- Transformera värdekedjor och sektorer till helt nya utvecklingsspår

4 Humancentrerad automation

Diskussioner om utformningen av AI-baserade system leder snabbt in på frågor som relaterar till den vetenskapliga litteraturen rörande humancentrerad automation, d.v.s MTO-frågor, där omfattande forskning och teoribildning finns att tillgå. Utformning av beslutsstöd och interaktion mellan människor och högt automatiserade system har diskuterats och utvecklats under lång tid, och även om det är inte rapportens syfte att återge detta i sin helhet ges nedan några exempel som är relevanta för AI diskussioner. För att ett AI-baserat system ska stödja en mänsklig operatör finns det en mängd frågor vars svar som måste vara tydliga:

- Perception – vad händer just nu?
- Notifikation – vad behöver jag veta?
- Situationsanpassning – vad behöver jag veta just nu?
- Automation – vad ska jag återkommande göra?
- Prediktion – vad kan jag förvänta mig ska ske?
- Prevention – vad kan jag undkomma?

WGHOF[48] *(Working Group on Human and Organisational Factors)* inom OECD lyfte 2012 fram dessa klassiska automationsfrågeställningar i sin identifiering av utmaningar med automation:

- Change in the overall role of operations and maintenance personnel
- Understanding the role of automation in operations
- Transition in workload for the operator when automation degrades or fails
- Monitoring of displays, vigilance or situation awareness, and complacency
- Out-of-the-loop unfamiliarity with underlying processes
- Degradation and loss of operational and maintenance skills
- The current trends in automation were described as:
    - New and advanced plants will be more highly automated
    - Expansion of the applications of automation
        - Greater use of automation for process control
        - Operator aids and decision support
        - HMIs that adjust without operator input
    - Greater range of the ways automation is implemented
        - Automation that is more interactive and cooperative
        - Shared control, breakpoint control (thresholds for changing functions among agents), dynamic allocation of functions among agents

Väl värda att beakta och ofta refererade är Bainbridges (1983) så kallade *ironies of automation*:

- The designers are not automated, and they are a major source of later errors and malfunctions
- The designer, who tries to eliminate the operator, still leaves the operator with the tasks which the designer cannot imagine how to automate

Den andra punkten ovan har av Hollnagel (1999) också kallats ”the left-over principle” där systemutvecklaren försöker automatisera operatörens arbetsuppgifter, men ändå lämnar operatören med de arbetsuppgifter som var för svåra att automatisera samt de som ansågs onödiga att automatisera ur ett kostnadsperspektiv. Arbetsuppgifterna däremellan, som ofta är de som gör att operatören vidmakthåller sin uppsikt och kontroll över systemet, automatiseras alltså bort från

48 http://www.oecd-nea.org/nsd/docs/2012/sen-sin-wghof2012-1.pdf

människan. Operatören får ta hand om resten och anses enligt denna princip därför vara i en sämre position att hantera avvikelser än när automationsprojektet påbörjades.

I sin artikel *Seven deadly myths of autonomous systems* beskriver Bradshaw, Hoffman, Johnson och Woods (2013) på generell nivå sina erfarenheter av automationsinförande, och jämför den faktiska komplexiteten som uppstod med de förväntade förbättringar som var anledningen till att en process/arbetsuppgift automatiserades. Tabell 1 nedan kondenserar många insikter, men återges här av utrymmesskäl utan vidare förklaringar, varvid de kursiverade texten kan upplevas svåra att förstå. Den intresserade läsaren hänvisas till orginalartikeln, men exempelvis avser *law of stretched systems* att även om automationen lyckas avlasta arbetsuppgifter från en operatör så kommer operatören själv, eller ledningen, att lägga till nya arbetsuppgifter tills operatörens kapacitet återigen ligger på fullt kapacitetsutnyttjande.

Tabell 1. Förväntade fördelar av autiomationsinförande jämfört med den resulterande komplexiteten (Bradshaw et al., 2013).

Expected benefit	**Real complexity**
Increased performance is obtained from "substitution" of machine activity for human activity.	Practice is transformed; the roles of people change; old and beloved habits and familiar features are altered – *the envisioned world problem*.
Frees up human by offloading work to the machine.	Creates new kinds of cognitive work for the human, often at the wrong times; every automation advance will be exploited to require people to do more, do it faster, or in more complex ways—*the law of stretched systems*.
Frees up limited attention by focusing someone on the correct answer.	Creates more threads to track; makes it harder for people to remain aware of and integrate all of the activities and changes around them—with coordination costs, continuously.
Less human knowledge is required.	New knowledge and skill demands are imposed on the human and the human might no longer have a sufficient context to make decisions, because they have been left out of the loop—*automation surprise*.
Agent will function autonomously.	Team play with people and other agents is critical to success—*principles of interdependence*.
Same feedback to human will be required.	New levels and types of feedback are needed to support peoples' new roles—with coordination costs, continuously.
Agent enables more flexibility to the system in a generic way.	Resulting explosion of features, options, and modes creates new demands, types of errors, and paths toward failure—*automation surprises*.
Human errors are reduced.	Machines, humans, and macro-cognitive work systems are fallible; errors are therefore systemic; new problems are associated with human-machine coordination breakdowns; machines now obscure information necessary for human decision making—*principles of complexity*.

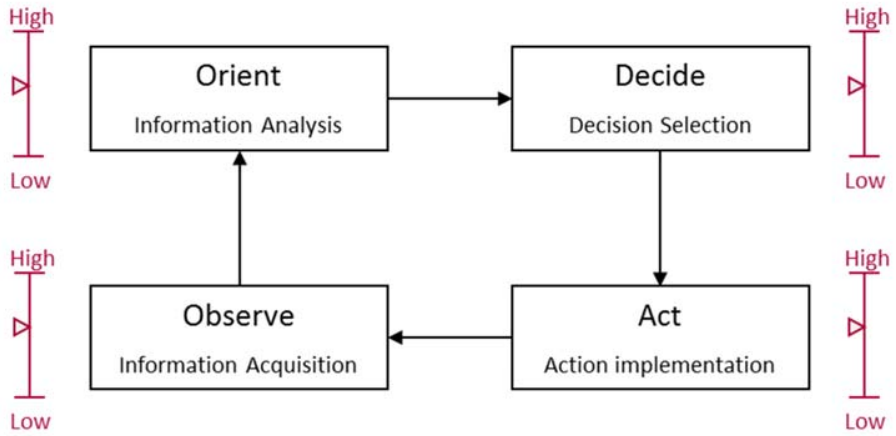
Många forskare inom den humancentrerade automationsforskningen har använt sig av så kallade *Levels of Automation* (LoA) skalor när de diskuterar automation. Den grundläggande idéen med LoA skalor är att det går att identifiera ett antal diskreta steg gällande mandatfördelningen från ett helt manuellt kontrollerat system upp till ett helt autonomt, maskinstyrt system. En av de tidigaste publicerade skalorna är Sheridan och Verplank (1978) som föreslår följande skala:

1. The computer offers no assistance, human must take all decisions and actions
2. The computer offers a complete set of decision/action alternatives, or
3. Narrows the selection down to a few, or
4. Suggests one alternative, and
5. Executes that suggestion if the human approves, or
6. Allows the human a restricted veto time before automatic execution
7. Executes automatically, then necessarily informs the human, and
8. Informs the human only if asked, or
9. Informs the human only if it, the computer, decides to
10. The computer decides everything, acts autonomously, ignores the human

Många alternativa skalor har publicerats och ett annat känt exempel är skalan av Billings (1997):

- Direct manual control: operator has direct authority over all systems
- Assisted manual control: operator is augmented by machine in his authority over systems
- Shared control: operator in control within envelope protection system
- Management by delegation: operator commands particular targets (e.g. heading and speed)
- Management by consent: operator must consent to system state changes
- Management by exception: operator monitors and intervenes in case of problems
- Autonomous operation: operator normally has no reason to intervene, even during anomalies

Parasuraman, Sheridan och Wickens (2000) fördjupade diskussionen genom att föreslå en modell där ett system kan ha olika automationsnivå för olika delfunktioner av systemet, se Figur 22. Deras modell grundar sig på en fyrdelad uppdelning av människans informationsprocessande och motsvarande systemfunktioner (i lådorna) och att systemet kan ha olika automationsnivå för dessa (skalorna *low-high*).



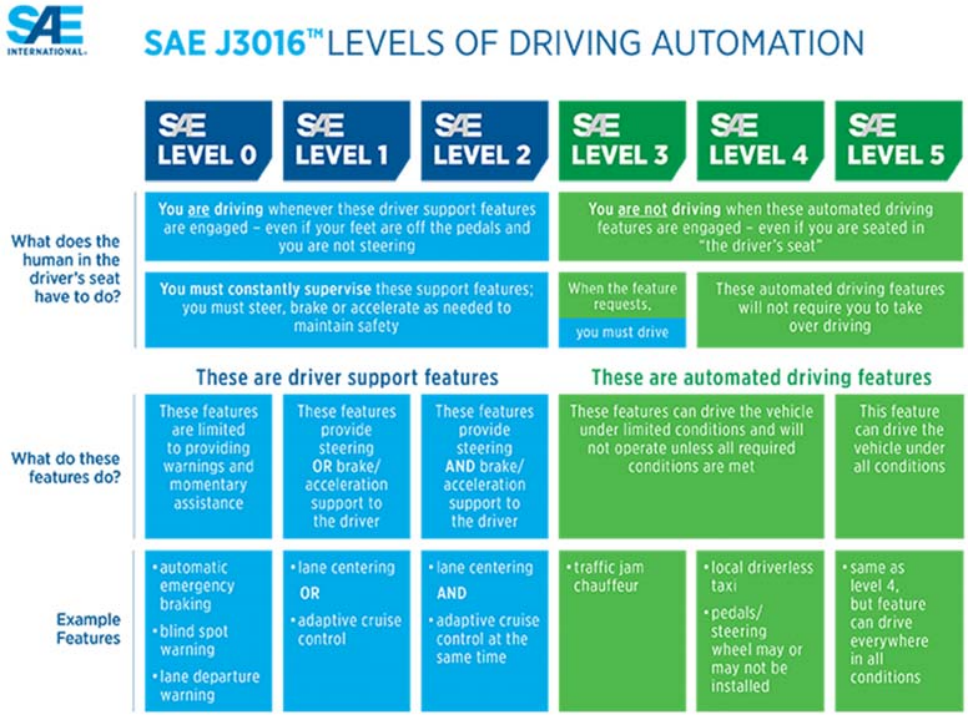
Figur 22. Teoretisk modell av automationsnivåer (Parasuraman, Sheridan och Wickens, 2000).

Ett aktuellt och mycket använt LoA exempel från utveckling av självkörande fordon är SAE (*Society of Automotive Engineers)* LoA-skala. Skalan återges i Figur 23 och Figur 24 i två versioner.

SAE level	Name	Narrative Definition	Execution of Steering and Acceleration/ Deceleration	*Monitoring* of Driving Environment	Fallback Performance of *Dynamic Driving Task*	System Capability *(Driving Modes)*
***Human driver* monitors the driving environment**						
0	**No Automation**	the full-time performance by the *human driver* of all aspects of the *dynamic driving task*, even when enhanced by warning or intervention systems	Human driver	Human driver	Human driver	n/a
1	**Driver Assistance**	the *driving mode*-specific execution by a driver assistance system of either steering or acceleration/deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	Human driver and system	Human driver	Human driver	Some driving modes
2	**Partial Automation**	the *driving mode*-specific execution by one or more driver assistance systems of both steering and acceleration/ deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	**System**	Human driver	Human driver	Some driving modes
***Automated driving system* ("system") monitors the driving environment**						
3	**Conditional Automation**	the *driving mode*-specific performance by an *automated driving system* of all aspects of the dynamic driving task with the expectation that the *human driver* will respond appropriately to a *request to intervene*	System	**System**	Human driver	Some driving modes
4	**High Automation**	the *driving mode*-specific performance by an automated driving system of all aspects of the *dynamic driving task*, even if a *human driver* does not respond appropriately to a *request to intervene*	System	System	**System**	Some driving modes
5	**Full Automation**	the full-time performance by an *automated driving system* of all aspects of the *dynamic driving task* under all roadway and environmental conditions that can be managed by a *human driver*	System	System	System	**All driving modes**

Figur 23. SAE:s LoA skala från 2014[49].

49 https://saemobilus.sae.org/content/j3016_201609



Figur 24. SAE:s LoA skala från 2018[50]*.*

50 https://www.sae.org/news/press-room/2018/12/sae-international-releases-updated-visual-chart-for-its-%E2%80%9Clevels-of-driving-automation%E2%80%9D-standard-for-self-driving-vehicles

Många forskare är kritiska mot användandet av LoA-skalor och anser att det snedvrider diskussioner och analyser. Johnson (2014) är en av dem och beskriver det istället som ett antal beroenden som måste kartläggas, se Tabell 2 nedan. Johnson beskriver sina observationer som gällande människa-robot samarbete, men de är tydligt överförbara till samarbete mellan en människa och ett AI-baserat system.

Tabell 2. Människa-robot interaktionsbehov (Johnson, 2014).

Human Needs	**Issues**	**Robot Needs**
What is the robot doing?	Mutual Transparency	What is the intent of the human?
Why did the robot do that?	Mutual 'Explainability'	What is the task context?
What is the robot going to do next?	Mutual Predictability	What does the human need from me?
How can we get the robot to do what we need?	Mutual 'Directability'	How can the human provide help?
Does use of autonomy add value?	Mutual Cost Benefit Management	Will my actions provide value to the human?

Klein m.fl. (2004) hävdar att en lagspelare, oavsett om det är en människa eller en maskin, måste kunna observera, förstå och förutsäga tillstånd och handlingar hos andra aktörer. Det finns många exempel på olyckor inom flertalet domäner där problemet varit "kraftfull och tyst automation", d.v.s. system som inte effektivt ger de signaler som gör det möjligt för mänskliga operatörer att förutsäga, kontrollera, förstå, och förutse vad det automatiska systemet är i för tillstånd eller hur det kommer agera. Vid behov av vidare fördjupning, som kan utvecklas till kravställning på ett AI-system, så beskriver Klein m.fl. tio stycken grundläggande utmaningar för samarbete mellan människor och automatiserade system.

Hollnagel och Woods (2005) introducerade begreppet *joint cognitive system* (JCS) som enkelt uttryckt avser ett system där mänskliga operatörer, tekniska system, och den relevanta omgivningen bör betraktas som ett gemensamt kognitivt system när systemet ska analyseras. I JCS-boken tar Hollnagel och Woods upp ett antal generella krav som också kan användas som embryo till högnivå-kravställning för AI-baserade system:

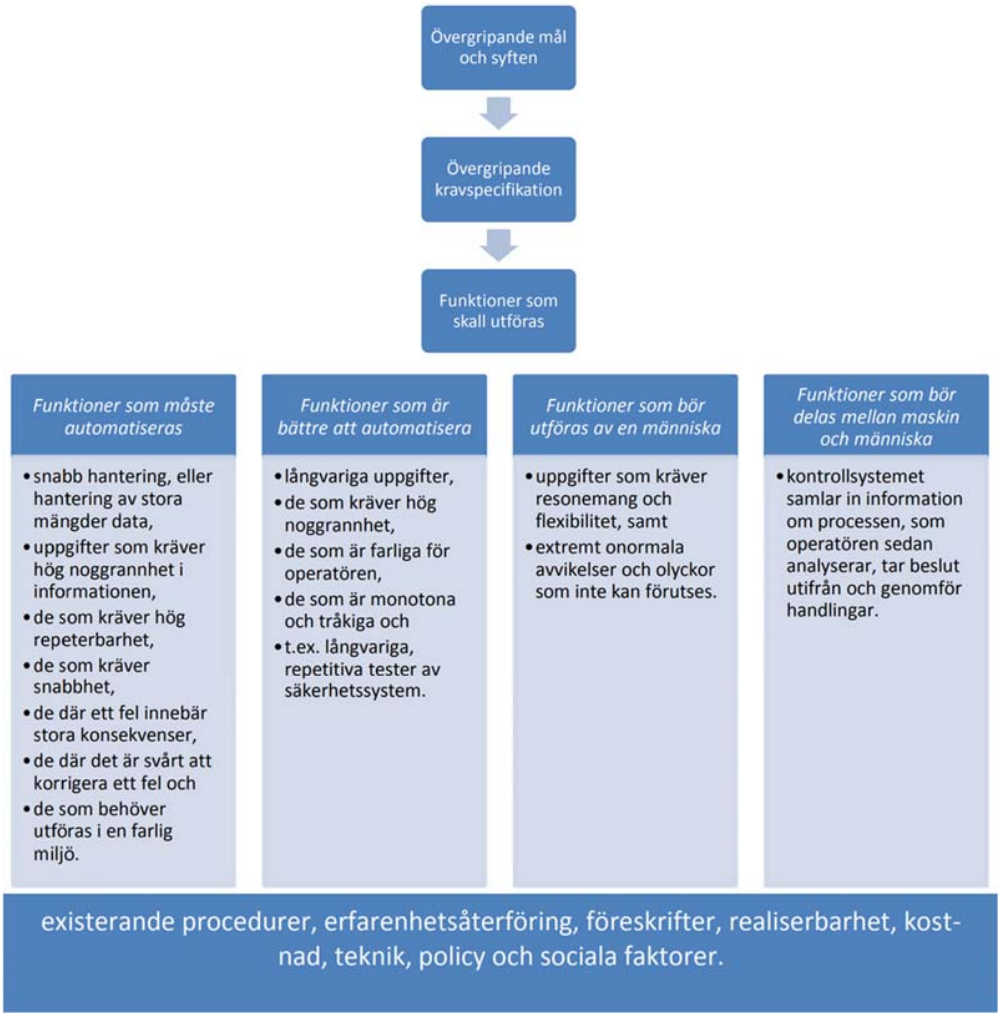
- **Support for Observability:** feedback that provides insight into a process
    - Integrate data based on a model of the process
    - Align data to reveal patterns and relationships in a process
    - Provide context around details of interest
    - Overcome "keyhole"/extend peripheral awareness
    - See sequence & evolution over time
    - See future activities & contingencies
    - Decompose integrations and inferences into sources, process, base evidence
- **Support for Directability:** ability to direct/re-direct resources, activities, priorities as situations change and escalate
    - Anticipation/projection
    - Models of capability
    - Policies for adaptation
    - Intent communication
- **Support for Directing Attention:** ability to re-orient focus in a changing world
    - Track others' focus of attention
    - Judge 'interruptibility' of others

    - Use Pre-attentive reference
- **Support for Shifting Perspectives:** contrasting points of view
    - Seeding—structure & kick start initial activity
    - Reminding—suggest other possibilities as activity progresses
    - Critiquing—point out alternatives as activities come to a close

Johnson (2014) beskriver sju myter gällande automation som är lika tillämpliga för diskussioner om AI:

- **Myth #1:** autonomy is unidimensional.
- **Myth #2:** the conceptualization of levels of autonomy is a useful scientific foundation for the development of autonomous system roadmaps.
- **Myth #3:** autonomy is a widget (i.e. a system function that can be inserted into an existing system without any major changes).
- **Myth #4:** autonomous systems are autonomous.
- **Myth #5**: once achieved, full autonomy obviates the need for human-machine collaboration.
- **Myth #6:** as machines acquire more autonomy, they will work as simple substitutes (or multipliers) of human capability.
- **Myth #7:** full autonomy is not only possible, but is always desirable.

SSM har tidigare beställt forskning rörande automation och hur den bör användas, exempelvis Lackman (2011) som beskriver ett antal fallstudier. Lackman har bland annat använt Parasuraman, Sheridan och Wickens (2000) ramverk för att beskriva processen rörande beslut om vad som kan/bör/måste/inte får automatiseras. Lackman beskriver processen för analys av automation enligt Figur 25.



Figur 25. Lackmans figur av processen vid automationsbeslut.

5 Organisatoriska, etiska och juridiska aspekter på AI

I det offentliga rummet pågår för närvarande en omfattande debatt om AI-införandets konsekvenser för en rad organisatoriska, etiska och juridiska frågor, som exempelvis hur många jobb som kommer påverkas[51] [52], om AI innebär slutet på mänskligheten samt vem som är ansvarig om det sker en olycka med ett AI-baserat system. Frågorna är givetvis viktiga att diskutera, men på samma sätt som många teknologiska innovationer så är det inte tekniken i sig utan hur människor väljer att använda dem som måste diskuteras, se exempelvis EU:s forskningskommissionär Carlos Moedas[53] uttala sig om AI på detta sätt (Moedas, 2017).

Erfarenheter från införandet av högt automatiserade system, exempelvis från den amerikanska försvarsmakten (DoD, 2012), visar att introduktionen av AI-baserade system ofta förändrar en verksamhet i grunden, vilket alla verksamheter bör ha beredskap för. En mängd frågor relaterande till etik och juridik kommer att behöva diskuteras.

Det finns flera sätt på vilket ett AI-baserat system kan få snedvridna beslutsgrunder eller respons på stimuli, exempelvis uppvisa rasistiska beslut i rekryteringsbedömningar, som är ett känt exempel från Amazon[54]. Ett annat uppmärksammat exempel med rasistiska kommentarer från en chat-bot är Tay[55] från Microsoft. Antingen programmeras snedvridna beslutsgrunder in direkt för att prioritera en viss typ av människor eller egenskap som ålder, utbildningsbakgrund, etnicitet eller kön före andra. En annan mer svårfångad snedvridning är när en mängd träningsdata, exempelvis från tidigare rekrytering eller tidigare Twitterinlägg används för att träna ett system som använder sig av neurala nätverk i någon form. Om exempelvis träningsdata visar att för ett visst urval ur en population så är risken för låg arbetsförmåga 10%, medan den för ett annat urval är 1%, så kommer systemet troligen favorisera rekrytering av det senare urvalet trots att 90% av medlemmarna i det första utvalet är lämpliga. Grundregler från statistiken, som exempelvis att korrelation inte nödvändigtvis implicerar kausalitet får inte glömmas bort, men ett neuralt nätverk har inte förmågan att förstå detta. Nätverket kan också klassificera på andra särdrag än de som utvecklarna sett framför sig.

Den viktiga poängen är att valet av träningsdata är avgörande, och om det finns andra historiska orsaker till att fördelningar i data ser ut som de gör så kommer detta finnas kvar i den klassificeringsalgoritm som systemet lär sig och använder framöver, vilket riskerar att bli ett självförstärkande system. Om ett processövervakningssystem tränas på en datamängd som innehåller en viss typ av olyckor och avvikelser så är det primärt denna typ av olyckor och avvikelser systemet kommer att känna igen. Det är därför centralt att förstå vem som har valt ut den datamängd systemet tränats med, vilka kriterier som användes för att välja data, ev. bortfallsanalyser om viss typ av data utelämnats, och hur systemet över tid tränas på nya data.

Nedan presenteras några utgångspunkter och frågor relaterande till olika typer av etiska och juridiska frågor som kan få relevans för SSM:s fortsatta analys av AI-utvecklingen. Detta stycke är delvis avsett att vara en referenskälla för SSM i det fall SSM får i uppgift att skriva riktlinjer för AI av organisatorisk,

51 https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf
52 https://www.mckinsey.com/~/media/mckinsey/featured%20insights/Digital%20Disruption/Harnessing%20automation%20for%20a%20future%20that%20works/MGI-A-future-that-works-Executive-summary.ashx
53 https://ec.europa.eu/commission/commissioners/2014-2019/moedas/announcements/stoa-annual-lecture-media-age-artificial-intelligence_en
54 https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
55 https://en.wikipedia.org/wiki/Tay_(bot)

etisk eller juridisk natur. Redundansen mellan de olika redovisningarna nedan är delvis hög och ger i flera fall rekommendationer där sammanhanget är hela samhället. De flesta av rekommendationerna nedan går dock vid behov att omforma till rekommendationer som skulle kunna vara användbara inom SSM verksamhetsområde.

5.1 EU JRC perspektiv

EU JRC (2018), EU Joint Research Centre, lyfter fram ett antal centrala perspektiv på AI-användning. Detaljbeskrivningarna är alltför omfattande för att inkludera i föreliggande rapport, men väl värda att använda som en utgångspunkt när en AI-satsning ska analyseras:

Ethical and societal perspective

- Individual level
  - Autonomy
  - Identity
  - Dignity
  - Privacy and data protection
- Societal level
  - Fairness and equity
  - Collective human identity and the 'good life'
  - Responsibility, accountability and transparency
  - Surveillance/datafication
  - Democracy and trust
  - Cumulative knowledge as a common good?

Legal perspective

- Transparency, explainability & liability
- Ownership, access and data sharing
- The protection of AI inventions / creations by intellectual property rights
- Regulatory approach

Educational perspective

- Direct AI impact on advanced skills demand
- Impact of AI on learning
  - Support current ways of working
  - Discover new ways of working
- Reduced demand for certain knowledge and skill sets
- Impact of AI on teaching

Economical perspective[56]

Cyber Security perspective

Computer processing and energy perspective

Data perspective

Societal Resilience perspective

56 För de sista fem perspektiven presenterar JRC tänkvärda analyser, men inget i kondenserad form som är lämpligt för denna rapport.

5.2 EU HLEG AI perspektiv

EU:s high-level expert group on AI presenterade i april 2019 sin rapport rörande "Trustworthy AI" (EU HLEG AI, 2019). Hundratals europeiska forskare med AI-expertis har varit inblandade i framtagandet av rapporten. Rapporten kan anses var en lämplig aktuell utgångspunkt för SMM:s i eventuellt arbete som exempelvis berör översyn och föreskrifter rörande verksamhet där AI-baserade system ingår. Rapporten har tre huvudsakliga delar.

I det första delen av rapporten beskrivs fyra grundläggande etiska principer för AI-utveckling. De fyra principerna som tas upp är:

- Respect for human autonomy
- Prevention of harm
- Fairness
- Explicability

I avsnitt två av rapporten beskrivs sju grundläggande krav som bör övervakas under hela AI-systemets livslängd:

- Human agency and oversight, including fundamental rights, human agency and human oversight
- Technical robustness and safety, including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility
- Privacy and data governance, including respect for privacy, quality and integrity of data, and access to data
- Transparency, including traceability, explainability and communication
- Diversity, non-discrimination and fairness, including avoidance of unfair bias, accessibility and universal design, and stakeholder participation
- Environmental and societal well-being, including sustainability and environmental friendliness, social impact, society and democracy
- Accountability, including auditability, minimisation and reporting of negative impact, trade-offs and redress

I sin tredje del innehåller rapporten en omfattande checklista med frågor, avsedd att vara användbar för olika nivåer och roller i en organisation. Listan är utformad för att vara generellt applicerbar över många typer av AI-system, varav alla inte är relevanta för SSM:s intresseområde. Checklistan innehåller dock många viktiga frågor som skulle kunna ingå i en SSM analys av ett AI-system, efter viss gallring. Listan innehåller likt övriga checklistor inga rekommendationer, utan innehåller enbart frågor som EU HLEG AI gruppen bedömt vara relevanta. Listan beskrivs som en första version som kommer utvärderas på flera ställen i Europa under cirka ett år. Hela checklistan med viktiga frågor från EU HLEG AI återges nedan:

TRUSTWORTHY AI ASSESSMENT LIST (PILOT VERSION) by EU HLEG AI (2019

1. Human agency and oversight

Fundamental rights:

- Did you carry out a fundamental rights impact assessment where there could be a negative impact on fundamental rights? Did you identify and document potential trade-offs made between the different principles and rights?

- Does the AI system interact with decisions by human (end) users (e.g. recommended actions or decisions to take, presenting of options)?
    - Could the AI system affect human autonomy by interfering with the (end) user's decision-making process in an unintended way?
    - Did you consider whether the AI system should communicate to (end) users that a decision, content, advice or outcome is the result of an algorithmic decision?
    - In case of a chat bot or other conversational system, are the human end users made aware that they are interacting with a non-human agent?

Human agency:

- Is the AI system implemented in work and labour process? If so, did you consider the task allocation between the AI system and humans for meaningful interactions and appropriate human oversight and control?
    - Does the AI system enhance or augment human capabilities?
    - Did you take safeguards to prevent overconfidence in or overreliance on the AI system for work processes?

Human oversight:

- Did you consider the appropriate level of human control for the particular AI system and use case?
    - Can you describe the level of human control or involvement?
    - Who is the "human in control" and what are the moments or tools for human intervention?
    - Did you put in place mechanisms and measures to ensure human control or oversight?
    - Did you take any measures to enable audit and to remedy issues related to governing AI autonomy?

- Is there is a self-learning or autonomous AI system or use case? If so, did you put in place more specific mechanisms of control and oversight?
    - Which detection and response mechanisms did you establish to assess whether something could go wrong?
    - Did you ensure a stop button or procedure to safely abort an operation where needed? Does this procedure abort the process entirely, in part, or delegate control to a human?

2. Technical robustness and safety

Resilience to attack and security:

- Did you assess potential forms of attacks to which the AI system could be vulnerable?
    - Did you consider different types and natures of vulnerabilities, such as data pollution, physical infrastructure, cyber-attacks?
- Did you put measures or systems in place to ensure the integrity and resilience of the AI system against potential attacks?
- Did you verify how your system behaves in unexpected situations and environments?
- Did you consider to what degree your system could be dual-use? If so, did you take suitable preventative measures against this case (including for instance not publishing the research or deploying the system)?

Fallback plan and general safety:

- Did you ensure that your system has a sufficient fallback plan if it encounters adversarial attacks or other unexpected situations (for example technical switching procedures or asking for a human operator before proceeding)?
- Did you consider the level of risk raised by the AI system in this specific use case?
    - Did you provide the necessary information in case of a risk for human physical integrity?
    - Did you consider an insurance policy to deal with potential damage from the AI system?
    - Did you identify potential safety risks of (other) foreseeable uses of the technology, including accidental or malicious misuse? Is there a plan to mitigate or manage these risks?
- Did you assess whether there is a probable chance that the AI system may cause damage or harm to users or third parties? Did you assess the likelihood, potential damage, impacted audience and severity?
    - Did you consider the liability and consumer protection rules, and take them into account?
    - Did you consider the potential impact or safety risk to the environment or to animals?
    - Did your risk analysis include whether security or network problems such as cybersecurity hazards could pose safety risks or damage due to unintentional behaviour of the AI system?
- Did you estimate the likely impact of a failure of your AI system when it provides wrong results, becomes unavailable, or provides societally unacceptable results (for example discrimination)?

- Did you define thresholds and did you put governance procedures in place to trigger alternative/fallback plans?
- Did you define and test fallback plans?

Accuracy:

- Did you assess what level and definition of accuracy would be required in the context of the AI system and use case?
  - Did you assess how accuracy is measured and assured?
  - Did you put in place measures to ensure that the data used is comprehensive and up to date?
  - Did you put in place measures in place to assess whether there is a need for additional data, for example to improve accuracy or to eliminate bias?
- Did you verify what harm would be caused if the AI system makes inaccurate predictions?
- Did you put in place ways to measure whether your system is making an unacceptable amount of inaccurate predictions?
- Did you put in place a series of steps to increase the system's accuracy?

Reliability and reproducibility:

- Did you put in place a strategy to monitor and test if the AI system is meeting the goals, purposes and intended applications?
  - Did you test whether specific contexts or particular conditions need to be taken into account to ensure reproducibility?
  - Did you put in place verification methods to measure and ensure different aspects of the system's reliability and reproducibility?
  - Did you put in place processes to describe when an AI system fails in certain types of settings?
  - Did you clearly document and operationalise these processes for the testing and verification of the reliability of AI systems?
  - Did you establish mechanisms of communication to assure (end-)users of the system’s reliability?

3. Privacy and data governance

Respect for privacy and data Protection:

- Depending on the use case, did you establish a mechanism allowing others to flag issues related to privacy or data protection in the AI system's processes of data collection (for training and operation) and data processing?
- Did you assess the type and scope of data in your data sets (for example whether they contain personal data)?
- Did you consider ways to develop the AI system or train the model without or with minimal use of potentially sensitive or personal data?
- Did you build in mechanisms for notice and control over personal data depending on the use case (such as valid consent and possibility to revoke, when applicable)?
- Did you take measures to enhance privacy, such as via encryption, anonymisation and aggregation?
- Where a Data Privacy Officer (DPO) exists, did you involve this person at an early stage in the process?

Quality and integrity of data:

- Did you align your system with relevant standards (for example ISO, IEEE) or widely adopted protocols for daily data management and governance?
- Did you establish oversight mechanisms for data collection, storage, processing and use?
- Did you assess the extent to which you are in control of the quality of the external data sources used?
- Did you put in place processes to ensure the quality and integrity of your data? Did you consider other processes? How are you verifying that your data sets have not been compromised or hacked?

Access to data:

- What protocols, processes and procedures did you follow to manage and ensure proper data governance?
    - Did you assess who can access users' data, and under what circumstances?
    - Did you ensure that these persons are qualified and required to access the data, and that they have the necessary competences to understand the details of data protection policy?
    - Did you ensure an oversight mechanism to log when, where, how, by whom and for what purpose data was accessed?

4. Transparency

Traceability:

- Did you establish measures that can ensure traceability? This could entail documenting the following methods:
    - Methods used for designing and developing the algorithmic system:
        - Rule-based AI systems: the method of programming or how the model was built
        - Learning-based AI systems; the method of training the algorithm, including which input data was gathered and selected, and how this occurred.
    - Methods used to test and validate the algorithmic system:
        - Rule-based AI systems; the scenarios or cases used in order to test and validate
        - Learning-based model: information about the data used to test and validate.
    - Outcomes of the algorithmic system:
        - The outcomes of or decisions taken by the algorithm, as well as potential other decisions that would result from different cases (for example, for other subgroups of users).

Explainability:

- Did you assess:
    - to what extent the decisions and hence the outcome made by the AI system can be understood?
    - to what degree the system's decision influences the organisation's decision-making processes?
    - why this particular system was deployed in this specific area?
    - what the system's business model is (for example, how does it create value for the organisation)?
- Did you ensure an explanation as to why the system took a certain choice resulting in a certain outcome that all users can understand?
- Did you design the AI system with interpretability in mind from the start?
    - Did you research and try to use the simplest and most interpretable model possible for the application in question?
    - Did you assess whether you can analyse your training and testing data? Can you change and update this over time?
    - Did you assess whether you can examine interpretability after the model's training and development, or whether you have access to the internal workflow of the model?

Communication:

- Did you communicate to (end-)users – through a disclaimer or any other means – that they are interacting with an AI system and not with another human? Did you label your AI system as such?
- Did you establish mechanisms to inform (end-)users on the reasons and criteria behind the AI system's outcomes?
    - Did you communicate this clearly and intelligibly to the intended audience?
    - Did you establish processes that consider users' feedback and use this to adapt the system?
    - Did you communicate around potential or perceived risks, such as bias?
    - Depending on the use case, did you consider communication and transparency towards other audiences, third parties or the general public?
- Did you clarify the purpose of the AI system and who or what may benefit from the product/service?
    - Did you specify usage scenarios for the product and clearly communicate these to ensure that it is understandable and appropriate for the intended audience?
    - Depending on the use case, did you think about human psychology and potential limitations, such as risk of confusion, confirmation bias or cognitive fatigue?
- Did you clearly communicate characteristics, limitations and potential shortcomings of the AI system?
    - In case of the system's development: to whoever is deploying it into a product or service?
    - In case of the system's deployment: to the (end-)user or consumer?

5. Diversity, non-discrimination and fairness

Unfair bias avoidance:

- Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design?
    - Did you assess and acknowledge the possible limitations stemming from the composition of the used data sets?
    - Did you consider diversity and representativeness of users in the data? Did you test for specific populations or problematic use cases?
    - Did you research and use available technical tools to improve your understanding of the data, model and performance?

    - Did you put in place processes to test and monitor for potential biases during the development, deployment and use phase of the system?

- Depending on the use case, did you ensure a mechanism that allows others to flag issues related to bias, discrimination or poor performance of the AI system?
    - Did you establish clear steps and ways of communicating on how and to whom such issues can be raised?
    - Did you consider others, potentially indirectly affected by the AI system, in addition to the (end)-users?
- Did you assess whether there is any possible decision variability that can occur under the same conditions?
    - If so, did you consider what the possible causes of this could be?
    - In case of variability, did you establish a measurement or assessment mechanism of the potential impact of such variability on fundamental rights?
- Did you ensure an adequate working definition of "fairness" that you apply in designing AI systems?
    - Is your definition commonly used? Did you consider other definitions before choosing this one?
    - Did you ensure a quantitative analysis or metrics to measure and test the applied definition of fairness?
    - Did you establish mechanisms to ensure fairness in your AI systems? Did you consider other potential mechanisms?

Accessibility and universal design:

- Did you ensure that the AI system accommodates a wide range of individual preferences and abilities?
    - Did you assess whether the AI system usable by those with special needs or disabilities or those at risk of exclusion? How was this designed into the system and how is it verified?
    - Did you ensure that information about the AI system is accessible also to users of assistive technologies?
    - Did you involve or consult this community during the development phase of the AI system?

- Did you take the impact of your AI system on the potential user audience into account?
    - Did you assess whether the team involved in building the AI system is representative of your target user audience? Is it representative of the wider population, considering also of other groups who might tangentially be impacted?

- Did you assess whether there could be persons or groups who might be disproportionately affected by negative implications?
- Did you get feedback from other teams or groups that represent different backgrounds and experiences?

Stakeholder participation:

- Did you consider a mechanism to include the participation of different stakeholders in the AI system's development and use?
- Did you pave the way for the introduction of the AI system in your organisation by informing and involving impacted workers and their representatives in advance?

6. Societal and environmental well-being

Sustainable and environmentally friendly AI:

- Did you establish mechanisms to measure the environmental impact of the AI system's development, deployment and use (for example the type of energy used by the data centres)?
- Did you ensure measures to reduce the environmental impact of your AI system's life cycle?

Social impact:

- In case the AI system interacts directly with humans:
  - Did you assess whether the AI system encourages humans to develop attachment and empathy towards the system?
  - Did you ensure that the AI system clearly signals that its social interaction is simulated and that it has no capacities of "understanding" and "feeling"?
- Did you ensure that the social impacts of the AI system are well understood? For example, did you assess whether there is a risk of job loss or de-skilling of the workforce? What steps have been taken to counteract such risks?

Society and democracy:

- Did you assess the broader societal impact of the AI system's use beyond the individual (end-)user, such as potentially indirectly affected stakeholders?

7. Accountability

Auditability:

- Did you establish mechanisms that facilitate the system's auditability, such as ensuring traceability and logging of the AI system's processes and outcomes?
- Did you ensure, in applications affecting fundamental rights (including safety-critical applications) that the AI system can be audited independently?

Minimising and reporting negative Impact:

- Did you carry out a risk or impact assessment of the AI system, which takes into account different stakeholders that are (in)directly affected?
- Did you provide training and education to help developing accountability practices?
    - Which workers or branches of the team are involved? Does it go beyond the development phase?
    - Do these trainings also teach the potential legal framework applicable to the AI system?
    - Did you consider establishing an 'ethical AI review board' or a similar mechanism to discuss overall accountability and ethics practices, including potentially unclear grey areas?
- Did you foresee any kind of external guidance or put in place auditing processes to oversee ethics and accountability, in addition to internal initiatives?
- Did you establish processes for third parties (e.g. suppliers, consumers, distributors/vendors) or workers to report potential vulnerabilities, risks or biases in the AI system?

Documenting trade-offs:

- Did you establish a mechanism to identify relevant interests and values implicated by the AI system and potential trade-offs between them?
- How do you decide on such trade-offs? Did you ensure that the trade-off decision was documented?

Ability to redress:

- Did you establish an adequate set of mechanisms that allows for redress in case of the occurrence of any harm or adverse impact?
- Did you put mechanisms in place both to provide information to (end-)users/third parties about opportunities for redress?

5.3 Asilomars AI-principer

Ett pågående initiativ är de så kallande Asilomars AI-principer[57], utgivna av Future of Life Institute med flera mycket kända deltagare. Definitionen av principerna och undertecknandet av tillhörande upprop har pågått sedan 2017. De 23 identifierade principerna beskriver på övergripande nivå vad som bör vara ingångsvärden för varje AI-satsning.

Research Issues

1. Research Goal: The goal of AI research should be to create not undirected intelligence, but beneficial intelligence.
2. Research Funding: Investments in AI should be accompanied by funding for research on ensuring its beneficial use, including thorny questions in computer science, economics, law, ethics, and social studies, such as:
   - How can we make future AI systems highly robust, so that they do what we want without malfunctioning or getting hacked?
   - How can we grow our prosperity through automation while maintaining people's resources and purpose?
   - How can we update our legal systems to be more fair and efficient, to keep pace with AI, and to manage the risks associated with AI?
   - What set of values should AI be aligned with, and what legal and ethical status should it have?
3. Science-Policy Link: There should be constructive and healthy exchange between AI researchers and policy-makers.
4. Research Culture: A culture of cooperation, trust, and transparency should be fostered among researchers and developers of AI.
5. Race Avoidance: Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards.

Ethics and Values

6. Safety: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.
7. Failure Transparency: If an AI system causes harm, it should be possible to ascertain why.
8. Judicial Transparency: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.
9. Responsibility: Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.
10. Value Alignment: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.
11. Human Values: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.
12. Personal Privacy: People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data.
13. Liberty and Privacy: The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.

57 https://futureoflife.org/ai-principles

14. Shared Benefit: AI technologies should benefit and empower as many people as possible.
15. Shared Prosperity: The economic prosperity created by AI should be shared broadly, to benefit all of humanity.
16. Human Control: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.
17. Non-subversion: The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.
18. AI Arms Race: An arms race in lethal autonomous weapons should be avoided.

Longer-term Issues

19. Capability Caution: There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities.
20. Importance: Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources.
21. Risks: Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact.
22. Recursive Self-Improvement: AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures.
23. Common Good: Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization.

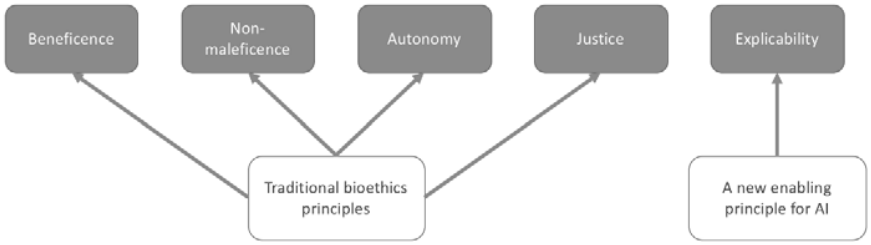
5.4 AlgoAware perspektiv

Det EU-finansierade AlgoAware-projektet resonerar i termer av system som styrs av algoritmer snarare än att explicit använda begreppet AI, men insikterna är tydligt giltiga för denna rapport. AlgoAware projektet lyfter fram följande viktiga aspekter att utvärdera och kravställa:

- Fairness and equity – in particular referring to the possible discriminatory results algorithmic decisions can lead to, and appropriate benchmarks automated systems should be assessed against.
- Transparency and scrutiny – algorithmic systems are complex and can make inferences based on large amounts of data where cause and effect are not intuitive. This concept relates to the potential oversight one might have on the systems.
- Accountability – a relational concept allowing stakeholders to interact, both to hold and to be held to account.
- Robustness and resilience – refers to the ability of an algorithmic system to continue operating the way it was intended to, in particular when re-purposed or re-used.
- Privacy – algorithmic systems can impact an individual's, or a group of individuals, right to private and family life and to the protection of their personal data; and
- Liability – questions of liability frequently arise in discussions about computational systems which have direct physical effects on the world (for instance self-driving cars). Tensions exist between some of these concepts. Ensuring the transparency of an algorithmic system might come at the expense of its resilience, whilst ensuring fairness may necessitate a relinquishing a degree of privacy.

5.5 Kondensat av etiska principer

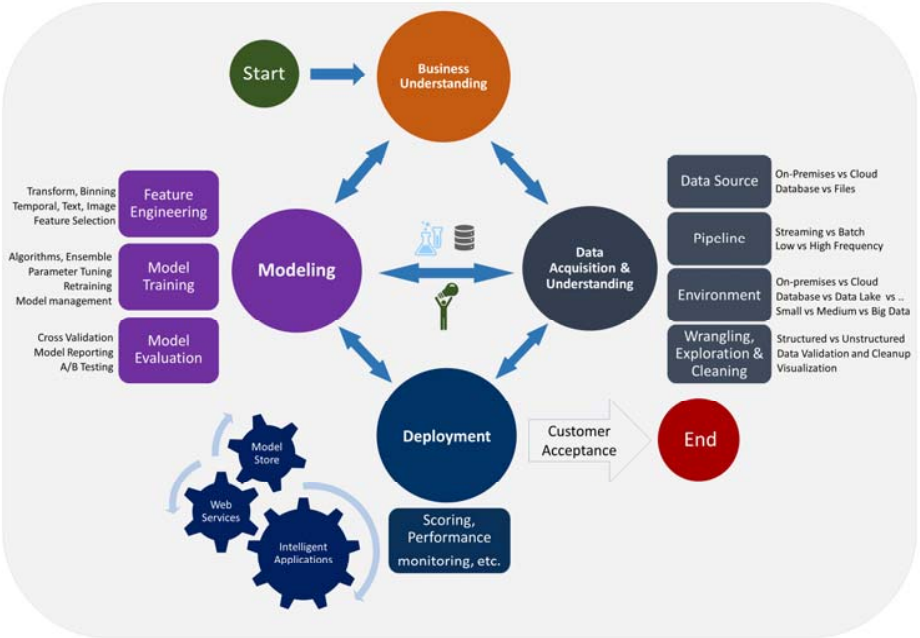
Floridi, Cowls, Beltrametti m.fl. (2018) lät en större grupp av experter inom AI, etik, och lagstiftning bedöma och kondensera 47 stycken AI och etikrelaterade principer som extraherats från sex större etiksammanställningar. Resultat blev ett förslag med fem centrala principer, se Figur 26, som bör beaktas vid utformning av etikdokument för AI. För ytterligare detaljer om innebörden av respektive princip hänvisas till Floridis m.fl. ursprungsartikel[58].



Figur 26. Floridis m.fl. (2018) etikprinciper för AI.

5.6 Utvecklingsprocess

Utvecklingsprocessen för AI-baserade system är knappast SSM:s ansvar, men i sammanhanget kan det ändå vara värt att betänkta att utvecklingsprocessen för ett AI-baserat system kan skilja sig från en normal systemutvecklingsprocess. Systemutvecklingsmetoden bör därför anpassas för aktuellt projekt. Ett exempel på en systemutvecklingsmetod som är anpassad för utvecklingen av intelligenta system är Microsofts Team Data Science Process[59], se Figur 27 för en schematisk beskrivning.



Figur 27. Team Data Science Process.

58 https://link.springer.com/article/10.1007/s11023-018-9482-5
59 https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview

6 Aktuella satsningar

Under denna rubrik presenteras ett axplock av aktuella AI-relaterade satsningar inom forskning, policyutveckling, och produktrelaterad AI-utveckling som bedöms vara relevanta för SSM fortsatta uppbyggnad av kompetens inom AI-området.

6.1 Amerikanska forskningsprojekt

6.1.1 MEITNER

ARPA-E (U.S. Department of Energy's Advanced Research Projects Administration-Energy) har under 2018 påbörjat finansiering av projektet *Development of a Nearly Autonomous Management and Control System for Advanced Reactors* under det så kallade MEITNER programmet. Professor Nam Dinh, som tidigare verkat vid KTH, leder projektet från North Carolina State University. Projektet finansieras 2018–2021 med ca 25 MKr. Projektet ska utveckla ett högt automatiserat system för styrning av avancerade kärnreaktorer. Genom omfattande övervakning samt simulering av anläggningens parametrar ska ett AI-baserat system användas för att förutsäga händelseutvecklingen och ge rekommendationer till kontrollrumsoperatörerna gällande åtgärder. Några länkar som översiktligt beskriver projektet är finns tillgängliga via fotnot[60] [61]. Resultat från projektet finns inte tillgängliga ännu, men projektet är ett av få tydliga kärnkraftsprojekt med AI-inriktning till vilket information hittats under författandet av rapporten.

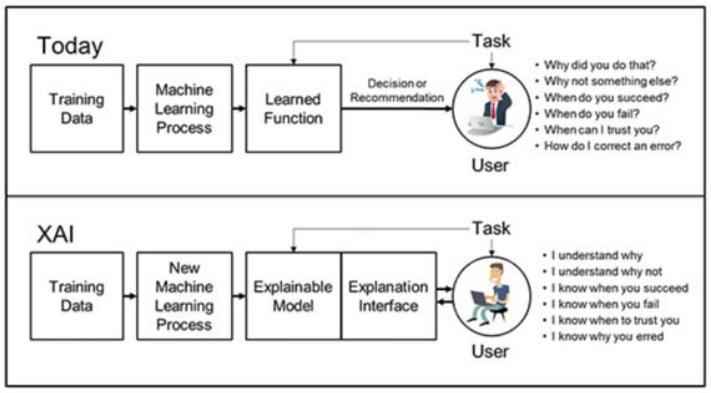
6.1.2 I4Gen

EPRI (Electric Power Research Institute) driver projekttet i4Gen som utvärderar användbarheten hos diverse prediktiva analystekniker[62] för prediktivt underhåll. Författarna har inte tillgång till artiklarna då detta kräver EPRI medlemskap, men EPRI verkar ha flera AI och maskininlärningsrelaterade initiativ i gång.

6.1.3 Explainable AI (XAI)

XAI[63] [64] är ett större forskningsprogram från amerikanska DARPA (*Defense Advanced Research Program Agency)*. DARPA finansierar de mest banbrytande teknikutvecklingsprojekten för den amerikanska försvarsmakten, med utvecklingen av internet och smygteknik som kända framgångsrika resultat. Behovet av förklaringsbar AI är givetvis inte nytt, utan har funnits hela tiden, men DARPA utlyste under 2016 efter ansökningar rörande XAI eftersom de ser detta som en av AI-områdets största utmaningar. Förklaringsbarheten och behovet av att förstå varför en AI-tillämpning agerar som den gör är givetvis inte nytt. Eftersom moderna AI-tillämpningar, särskilt de som använder subsymboliska ansatser, ofta är att betrakta som stängda lådor där användaren har mycket svårt att förstå varför resultat blir som de blir, ökar behovet att ett förklarande gränssnitt mellan människan och AI-algoritmerna, schematiskt beskrivet i Figur 28.

60 https://arpa-e.energy.gov/?q=slick-sheet-project/management-and-control-system-advanced-reactors
61 http://www.govtech.com/computing/AI-Controlling-Nuclear-Reactors-It-Could-Happen.html
62 https://www.epri.com/#/pages/product/000000003002011560/?lang=en-US
63 https://vimeo.com/237663041
64 https://www.darpa.mil/program/explainable-artificial-intelligence



Figur 28. XAI-programmets visualisering av programmets syfte.

De första resultaten från XAI-programmet har börjat bli tillgängliga, exempelvis Mueller, Hoffman, Klein m.fl. (2018) som beskriver hur dagens AI, särskilt lösningar som använder sig av olika neurala nätansatser är exempel på en ny sorts annorlunda intelligens, med en rationalitet som inte nödvändigtvis hittar de mönster som en människa lärt sig och förväntar sig. Mueller m.fl. (2018)[65] presenterar ett pedagogiskt exempel på hur nätverk tränas på att klassificera blommor, men sedan klassificerar en röd bil som en röd tulpan vid överföring till en annan datamängd. De särdrag ett neuralt nätverk använder sig av för att klassificera i en datamängd behöver inte alls vara de som människan antar att de använder och förståelsen för hur klassificering fungerar kan vara svår att bygga upp för en användare. Mueller m.fl. (2018) lyfter fram punkterna nedan som viktiga insikter rörande AI och specifikt neurala nätverk:

- AI is an Alien Intelligence
- Their behavior does not map onto our intuitions
- They can be good at fine discriminations but terrible at large ones
- They appear to focus on local features at the cost of global properties
- We cannot ask them, and they cannot tell us why they are acting strangely bad or incredibly well

Mueller, Hoffman, Clancey, Emrey och Klein (2019) presenterar en mycket omfattande litteraturöversikt med mer än 700 referenser rörande forskningen kring förklaring generellt och specifikt förklaring från AI-system. AI får inte bli system utan insyn, vars agerande inte går att förstå för en mänsklig operatör. Litteraturöversikten innehåller en mängd information för djupare studier. Tabell 3 ger ett exempel på identifiering av på centrala frågeställningar gällande förklaringar från ett AI-system.

Tabell 3. Mueller m.fl. (2019) exempel på centrala frågeställningar gällande förklaring från ett AI-system.

TRIGGERS	**USER GOAL**
How do I use it?	Achieve the primary ask goals
How does it work?	Feeling of satisfaction at having achieved an understanding of the system, in general (global understanding)

65 https://www.researchgate.net/publication/329352442_Mueller_S_T_2018_Panel_on_Explainable_Artificial_Intelligence_XAI_in_Proceedings_of_the_23rd_International_Command_and_Control_Research_Technology_Symposium_ICCRTS_6-9_Nov_Pensacola_FL_USA

What did it just do?	Feeling of satisfaction at having achieved an understanding of how the system made a particular decision (local understanding)
What does it achieve?	Understanding of the system's functions and uses
What will it do next?	Feeling of trust based on the observability and predictability of the system
How much effort will this take?	Feeling of effectiveness and achievement of the primary task goals
What do I do if it gets it wrong?	Desire to avoid mistakes
How do I avoid the failure modes?	Desire to mitigate errors
What would it have done if x were different?	Resolution of curiosity at having achieved an understanding of the system
Why didn't it do z?	Resolution of curiosity at having achieved an understanding of the local decision

6.2 Nationella AI agendor inom EU

I april 2018 undertecknade EU:s medlemsnationer ett samarbetsavtal kring AI-utveckling. Många av länderna har initierat utveckling av nationella AI agendor. Många av dessa agendor och presentationer av dem finns fritt tillgängliga[66] [67].

EU har också tillsatt en expertgrupp rörande AI *(HLEG AI, High Level Expert Group on Artificial Intelligence)* som ska ta fram riktlinjer gällande AI-relaterade etiska frågor samt beskriva AI-frågor som på medellång och lång sikt kommer påverka utvecklingen av riktlinjer och lagstiftning samt utvecklingen av EU:s nästa digitala strategi. Se EU HLEG AI 2018a[68], 2018b[69], 2019[70].

6.3 EU finansierade projekt

EU driver ett antal forskningsprojekt som relaterar till tekniska frågeställningar samt etik och policyfrågor kring AI. Fyra exempel presenteras nedan, och resultaten från dessa satsningar kan komma att bli embryon till diverse EU riktlinjer som kan vara relevanta för SSM.

6.3.1 RAIN

Robotics and Artificial Intelligence for Nuclear (RAIN)[71] [72] projektet är ett forskningsprojekt med 36 europeiska projektpartner och med åtta huvudsakliga akademiska forskningsutförare, lett från University of Manchester och finansierat av UK Research and Innovation. Fjärrstyrd övervakning och systemhantering med diverse robotikapplikationer verkar vara i fokus.

6.3.2 AI4EU

AI4EU[73] [74] är ett forskningsprojekt finansierat av EU, lett av Thales som påbörjas i januari 2019 och planeras pågå i tre år, finansierat med ca 180 MKr. 79 deltagande organisationer ska ta fram åtta pilotprojekt för att stärka europeisk samverkan inom AI-området och det finns en ambition om en

66 https://ec.europa.eu/digital-single-market/en/news/european-artificial-intelligence-landscape
67 https://futureoflife.org/national-international-ai-strategies
68 https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56018
69 https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december.pdf
70 https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
71 https://rainhub.org.uk/
72 https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/R026084/1
73 http://ai4eu.org
74 https://ec.europa.eu/digital-single-market/en/news/artificial-intelligence-ai4eu-project-launches-1-january-2019

öppen gemensam AI-plattform samt att projektet ska ge rekommendationer till EU rörande etik och lagstiftning inom området.

6.3.3 AlgoAware

AlgoAware[75] är ett forskningsprojekt finansierat av EU. I sin *state of the art* rapport (AlgoAware, 2018) från december 2018 ges en omfattande översikt över AI området inom Europa utifrån flera olika perspektiv, se bl.a. avsnitt 5.4.

6.3.4 European AI alliance

Med styrning från EU HLEG AI, har EU satt upp European AI Alliance[76], avsett att vara ett diskussionsforum för etik och policyrelaterade frågor rörande AI. För nuvarande är ca 2500 forskare och utvecklare från olika typer av organisationer anmälda som medlemmar. EU HLEG AI använder bland annat detta diskussionsforum för att få återkoppling på sina utkast till EU AI-riktlinjer. Ett exempel som publicerades i april 2019 är dokument *EU Ethics guidelines for trustworthy* AI (EU HLEG AI, 2019).

6.3.5 CLAIRE

CLAIRE[77] - *Confederation of Laboratories for Artificial Intelligence Research in Europe*, är ett nätverk för AI aktiva forskare i Europa, upprättat 2018, som strävar att stärka AI-forskningen i Europa. Mer än 1000 doktorer och professorer deltar i nätverket.

6.3.6 SIENNA

SIENNA-projektet *(Stakeholder-Informed Ethics for New technologies with high socio-ecoNomic and human rights impAct)*[78] är ett EU finansierat project som påbörjats under 2018 som planerar verksamhet som ska belysa etik inom AI området.

6.3.7 SHERPA

SHERPA-projektet[79] är ett EU finansierat projekt som pågår 2018–2021 som kommer belysa ett europeiskt perspektiv på etik och mänskliga rättigheter när smarta informationssystem används i ökande utsträckning.

6.3.8 PANELFIT

PANELFIT-projektet[80] [81] är ytterligare ett exempel på ett EU-finansierat projekt, ca 30 Mkr, som kommer studera etiska och legala utmaningar med nya IT-system.

6.4 Nordiska AI-satsningar

6.4.1 Uniper - OKG

Vid OKG finns en mycket stor datamängd tillgänglig, där data från huvuddelen av anläggningens sensorer finns registrerade på minut eller sekundnivå sedan 2010. OKG har ett pågående projekt, kontaktperson Sune Jonsson, påbörjat våren 2018, där AI och maskininlärningsalgoritmer kommer att användas för att analysera denna datamängd. Syftet är att utveckla ett system som använder maskininlärningens förmåga till mönsterigenkänning för att kunna identifiera trender och avvikelser i data och använda detta för att larma operatörer. Tillämpningar finns dock även gällande smart

75 https://www.algoaware.eu
76 https://ec.europa.eu/digital-single-market/en/european-ai-alliance
77 https://claire-ai.org
78 http://www.sienna-project.eu
79 https://www.project-sherpa.eu/understanding-ethics-and-human-rights-in-smart-information-systems
80 http://www.idala.es/bioderecho
81 https://cordis.europa.eu/project/rcn/218355/factsheet/en

underhåll, där trendinformation används som underlag för exempelvis ett filterbyte. Systemet bör också bli användbart för att stödja orsaksutredning *(root cause analysis)* i efterhand. Avsikten är att initialt under 2019 genomföra *proof of concept* analyser på en delmängd av data för utvalda delsystem, men sedan gå vidare till att analysera hela datamängden.

Kunskapsspridning rörande potentialen med AI sker både genom bilaterala möten mellan kärnkraftverken och i andra kanaler som exempelvis *System Sustainability Circle* [82]. Eftersom mängden produkter och studier specifikt rörande AI och kärnkraft är relativt begränsad utvärderas i förekommande fall tillämpningar och system utvecklade för annan processindustri. Givet kärnkraftens mycket höga säkerhetskrav är datasäkerhet hos leverantörer och system alltid högt prioriterat.

6.4.2 WASP

Den enskilt största aktuella, 2018 och några år framöver, AI-satsningen i Sverige är WASP programmet *(Wallenberg Autonomous Systems Program)*[83]. Med riklig finansiering från Wallenbergstiftelsen har en mängd AI-relaterade doktorandarbeten påbörjats. Det övergripande syftet med WASP är att studera vetenskapliga och praktiska utmaningar som gäller alla autonoma system, exempelvis behovet för system att lära sig och att samarbeta. Programmet berör maskininlärning, djupinlärning, och explainable AI samt underliggande teoretiska frågor rörande AI i bredare bemärkelse. Forskningen är utformad för att strategiskt gynna svensk industriell och akademisk utveckling inom flera domäner som sjukvård, finans, tjänsteutveckling, industriell utveckling samt samhällsutveckling. Programmet omfattar flera miljarder kronor och avser finansiera ca 100 doktorander, konferensserier, infrastruktur och karriärprogram.

6.4.3 MonitorX

MonitorX[84] är ett samarbetsprojekt som avslutas 2019, lett av EnergiNorge med deltagande från ca 35 svenska och norska aktörer (ungefärlig omfattning 20 MKr) med intresse för digitalisering inom vattenkraften. Inom projektet har flera aktörer med intresse för ökad förmåga till fjärrövervakning och förebyggande service beskrivit sina behov, där diverse AI-tekniker kan vara användbara och delvis motsvarar de som finns inom kärnkraften.

6.4.4 CHAIR

Chalmers AI Research Centre (CHAIR)[85] är ett kompetenscenter för AI vid Chalmers tekniska högskola som startats under 2019.

6.4.5 AI Innovation of Sweden

AI Innovation of Sweden[86], ett kompetenscenter inom AI, öppnade under 2019 och återfinns på Lindholmen i Göteborg. Fokus kommer att ligga på att accelerera tillämpningen av AI genom delning av kunskap och data, samlokalisering och samarbetsprojekt, allt med ett starkt fokus på etik, transparens och säkerhet. Satsningen finansieras av Vinnova (30 Mkr över 4 år) samt deltagande från ett 40-tal medlemmar som förväntas bidra med egen tid.

82 https://www.sustainabilitycircle.se/sustainability-circle-1
83 http://wasp-sweden.org
84 https://www.sintef.no/projectweb/monitorx
85 https://www.chalmers.se/en/centres/chair/Pages/default.aspx
86 https://www.ai.se/en

6.4.6 SAIS

Svenska AI sällskapet (SAIS) [87] har varit aktivt sedan 1982 och samlar forskare från många svenska akademiska lärosäten som är aktiva inom AI. Sällskapet kan vara lämpligt som ingång för att identifiera kontaktpersoner om någon granskningskommitté ska sättas upp. SAIS (2018) beskriver viktiga trender inom AI området som:

- Autonomous systems and Robotics
- Machine learning
- Explainable AI
- Data driven decision making
- Semantic technologies
- Natural language understanding
- Cognitive computing / Q&A-systems

SAIS listar också starka svenska forskningsområden:

- Knowledge representation and reasoning
- Planning
- AI/Robotics
- Constraint Programming
- Reinforcement Learning
- Deep Learning
- Bayesian Learning
- Data Mining
- Multi-agent systems
- Natural Language Processing
- Explainable AI
- Human-AI Interaction
- Semantic Web

6.5 Internationella företags produkter och arkitekturer

Flera multinationella företag som Google, Apple, Facebook, Amazon, och IBM utvecklar uppmärksammande AI-produkter och öppna arkitekturer, se exempel i listan nedan. Flera dessa öppna arkitekturer är de arkitekturer som både dagens och morgondagens mjukvaru-utvecklare bedöms komma utgå ifrån när de utvecklar applikationer åt kärnkraftsindustrin.

- TensorFlow[88] är en mycket brett använd, öppen AI-arkitektur från Google. Den används av många utvecklare som vill skapa egna AI-tillämpningar. Ett exempel är Googles egen Google Assistant/Duplex[89] som är en chat-bot som kan utföra uppgifter som bokning av resor eller restauranger.
- Facebook stöder utvecklingen av de öppna AI-biblioteken PyTorch[90] och Caffe2[91].

87 http://www.sais.se
88 https://www.tensorflow.org
89 https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html
90 https://pytorch.org
91 https://caffe2.ai

- Amazon har introducerat Alexa[92] som är en AI-baserad assistent som man interagerar med genom naturligt språk, avsedd för flera användningsområden som att styra det uppkopplade hemmet, beställa mat och liknande tjänster.
- Apple utvecklar Siri[93] som är en digital assistent som man interagerar med genom tal och som kan utföra enklare uppgifter och söka information. SIRI utvecklas också för att få bättre datorsyn. Apple utvecklar också Core ML[94] som är ett API *(Application Programming Interface)* genom vilket fortsatt utveckling av AI-baserade appar ska bli enklare.
- Watson[95] är en känd tillämpning från IBM som besegrade de främsta mänskliga mästarna i TV-frågesportsprogrammet Jeopardy 2011. Watson har sedan dess kommersialiserats för ett antal tillämpningar av ”fråga-svar” karaktär inom sjukvård, finans, lag och rätt, samt försäljning.
- General Electric, GE, satsar likt många andra stora företag på AI. Inom GE Research utvecklas AI-baserade tillämpningar för feldiagnos, prognoser, process automation och preskriptiv analys som kan användas för risk och säkerhetsbedömningar och beslutsstöd[96].
- Peltarion[97] är ett svenskt företag, långt ifrån lika stort som övriga företag i listan, men som utvecklar en uppmärksammad AI-arkitektur som ska göra det enklare att utveckla egna AI-tillämpningar och kunna utnyttja potentialen i maskininlärning.
- Exelon[98] utvecklar ett antal olika AI-baserade produkter som kan användas för underhållsanalys inom process och kärnkraftsindustri.
- OpenAI[99] är ett icke-vinstdrivande företag som attraherat några av AI-områdets mest namnkunniga forskare och som har som uttalat syfte att utveckla säker och ”människovänlig” generell artificiell intelligens (AGI). OpenAI publicerar normalt källkoden för sina uppmärksammade projekt, men valde nyligen att inte publicerade koden för sitt verktyg för AI-baserad textgenerering, GPT-2, då de såg att GPT-2 kunde generera längre texter där det var alltför svårt att avgöra att de var AI-genererade.

6.6 ISO

Standardiseringsorganisationen ISO har flera aktiva arbetsgrupper rörande AI, benämnda ISO/IEC JTC 1/SC 42. På ISO:s websida[100] presenteras nuvarande läge gällande standardarbete för bl.a. terminologi, användningsfall, trovärdighet, och implikationer för styrning av AI i verksamheter. Inriktningen för de olika arbetsgrupperna återges nedan:

- **SC42/WG1 – The foundational standards working group**, which will take on two ongoing standardisation projects: Artificial Intelligence Concepts and Terminology ISO/IEC AWI 22989 and Framework for Artificial Intelligence Systems Using Machine Learning ISO/IEC AWI 25053.
- **SC42/SG1 – The computational approaches and characteristics of artificial intelligence study group**, which will study:
    - different technologies (including machine learning algorithms) used by AI systems including their characteristics and properties.

92 https://en.wikipedia.org/wiki/Amazon_Alexa
93 https://en.wikipedia.org/wiki/Siri
94 https://developer.apple.com/machine-learning
95 https://www.ibm.com/watson
96 https://www.ge.com/research/technology-domains/artificial-intelligence/machine-learning
97 https://peltarion.com/platform
98 https://energiforskmedia.blob.core.windows.net/media/24312/11-exelon_ansley.pdf
99 https://openai.com
100 https://www.iso.org/committee/6794475/x/catalogue/p/0/u/1/w/0/d/0

    - existing specialised AI systems, such as natural-language processing or computer vision systems, with the objective of understanding their computational architectures and approaches.
    - industry practices, processes and methods for the application of AI systems.
- **SC42/SG2 – The trustworthiness study group**, which will:
    - Investigate approaches to establish trust in AI systems through transparency, verifiability, explainability, controllability, etc.
    - Investigate engineering pitfalls and assess typical associated threats and risks to AI systems with their mitigation techniques and methods.
    - Investigate approaches to achieve AI systems' robustness, resiliency, reliability, accuracy, safety, security, privacy, etc.
    - Investigate types of sources of bias in AI systems with a goal of minimization, including but not limited to statistical bias in AI systems and AI aided decisionmaking.
- **SC42/SG3 – The use cases and applications study group**, which will:
    - Identify different AI application domains (e.g., social networks and embedded systems) and the different context of their use (e.g., fintech, health care, smart home, and autonomous cars).
    - Collect representative use cases.
    - Describe applications and use cases using the terminology and concepts defined in ISO/IEC AWI 22989 and ISO/IEC AWI 23053 and extend the terms as necessary.
    - Develop new work item proposals as appropriate and recommend placement.

7 Framåtblick

AI-området utvecklas mycket kraftigt och fler produkter, AI-arkitekturer, och specifika algoritmer för att lösa nödvändiga beräkningsproblem är att förvänta. För SSM:s tillsyn kommer många av frågorna relaterade till de humancentrerade automationsfrågeställningarna fortsatt att vara aktuella under lång tid, se avsnitt 4, med ständigt återkommande frågor kring när system får eller måste agera själv och hur operatörerna ska interagera med och få förtroende för systemet. Historiskt sett har det förekommit ett antal större forskningsprojekt rörande AI, oftast amerikanska försvarforskningsprojekt med mycket höga ställda ambitioner, som misslyckats med att infria förväntningarna. Detta har lett till kraftigt minskade anslag vilket utlöst så kallade AI-vintrar, när forskningen har haft svårt med finansieringen.

Den senaste vågen av genombrott för AI-området som helhet och diverse maskinlärningsansatser i synnerhet, har dock förflyttat AI från forskning till praktisk tillämpning i stor skala. Det finns just nu inget som antyder att utvecklingen kommer avstanna och AI kan därmed förväntas få en stor, genomgripande påverkan på många mänskliga verksamhetsområden. En viktig insikt rörande AI är också att till skillnad från andra IT relaterade genombrott, exempelvis övergången till PC datorer, introduktionen av Internet, och introduktionen av mobila plattformar, som alla haft en tydlig koppling till hårdvaruutveckling, så sker AI-utveckling nästan uteslutande i mjukvara, vilket möjliggör en mycket högre utvecklingshastighet. AI bör definitivt betraktas som en så kallad disruptiv teknologi, som förändrar vissa grundläggande förutsättningar för hur en verksamhet genomförs. EU (2017) beskriver det som en av århundradets mest strategiska teknologier. Rapportförfattarna håller med i denna beskrivning, men påminner också om den så kallade Amaras lag[101], som säger *We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run*.

Armstrong, Sotala och ÓhÉigeartaigh (2014) beskriver hur teknikspaning är svårt, särskilt för AI. De jämförde 95 olika prediktioner gjorda mellan 1950 och 2012 som AI-experter gjort för när AGI, Artificiell Generell Intelligens, kommer få sitt riktiga genomslag. De flesta av dessa prediktioner beskrev hur denna tidpunkt bedömdes ligga inom 20 år från när prediktionen gjordes. Mullins (2012) analys av mer än 2000 teknikförutsägelser från 300 tekniskspaningssammanställningar handlar inte bara om AI, utan är en mer generell analys, men beskriver hur teknikförutsägelser som ligger mer än tio år bort inte har större prediktionssäkerhet än att singla slant.

Ett exempel på en prediktion, dock fristående från analysen av Armstrong m.fl., kan vara ett citat ur Jastrows (1982) artikel om den tänkande datorn:

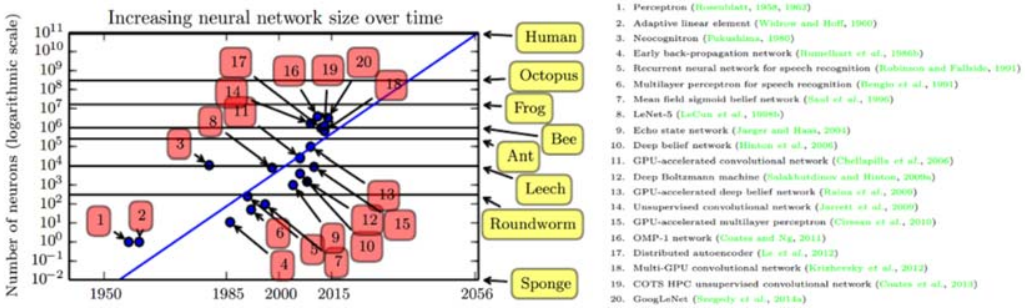
> *In five or six years - by 1988 or thereabouts - portable, quasi-human brains, made of silicon or gallium arsenide, will be commonplace. They will be an intelligent electronic race, working as partners with the human race. We will carry these small creatures with us everywhere. Just pick them up, tuck them under your arm, and go off to attend your business. They will be R2D2's without wheels: brilliant but nice personalities, never sarcastic, always giving you a straight answer – little electronic friends that can solve all your problems.*

Om man analyserar Jastrows prediktion så har vi idag kraftfulla verktyg i fickan genom våra mobiltelefoner, men att kalla dem en ny intelligent ras känns relativt avlägset för de flesta. Ett annat exempel, dock inte rörande AI, kan vara introduktionen av internet. Internet har definitivt förändrat många verksamheter genom att ge människor nya kanaler att lära sig saker, arbeta på, och ha sociala

101 https://en.wikipedia.org/wiki/Roy_Amara

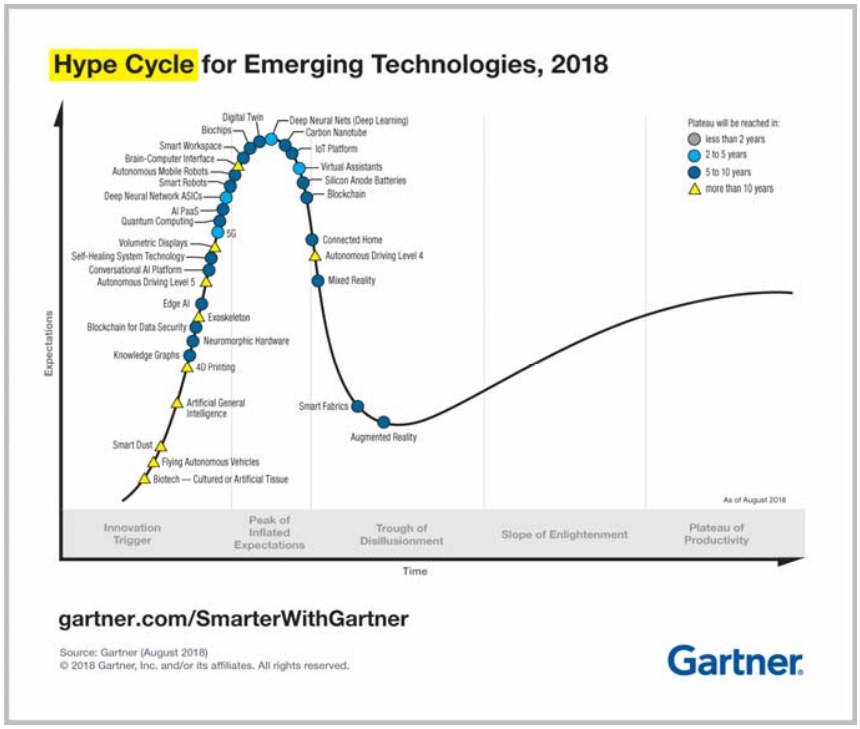
kontakter genom samt skapat nya produkter, tjänster och affärsmodeller. Prediktionerna om städernas död eller att allt arbete görs på distans har dock ännu inte infriats.

En intressant jämförelse kan vara den mellan storleken på de neurala nätverk som finns i ett antal olika biologiska system med storleken på ett antal artificiella neurala nätverk, se Figur 29. Figuren är från Goodfellow m.fl. (2016), som i sin tur återanvänt en bild från Wikipedia, och ger en uppfattning hur antalet neuroner som ingår i ANN har utvecklats under åren och deras relation till biologiska varelsers motsvarigheter. De röda siffrorna från 1–20 refererar till ett antal exempel på ANN ur forskningslitteraturen. Observera dock att det inte bara är antalet neuroner som avgör en hjärnas förmåga att hantera information. I en biologisk hjärna finns det ett antal fler komplexa interaktioner av till exempel neurotransmittorer i varje cell som avgör när den aktiveras. Goodfellows extrapolering till år 2056 som tidpunkten för när ANN har samma omfattning som den mänskliga hjärnan behöver därför inte vara helt relevant.



Figur 29. Antalet neuroner i ANN jämfört med biologiska hjärnor (ur Goodfellow m.fl., 2016).

Företaget GARTNER publicerar årligen sin analys av *the hype cycle*[102] för framväxande innovationer, se Figur 30. Avseende kurvan för år 2018 så är det värt att notera hur djupinlärning/*deep learning* befinner sig allra högst upp på toppen av förväntningskurvan.



Figur 30. Gartners Hype Cycle for Emerging Technologies 2018.

VINNOVA anger i sin rapport ”Artificiell intelligens i svenskt näringsliv och samhälle - Analys av utveckling och potential” (2018)[103] ett antal generella utmaningar för svensk industri som bedöms bli konsekvenser av ökad AI-användning. Följande faktorer och samspelen mellan dessa bedöms bli viktiga i alla verksamheter:

- **Affärs- och verksamhetsmodeller:** för vissa företag och offentliga verksamheter är den värdeskapande potentialen med AI tydlig, medan andra ännu inte ser nyttan lika tydligt.
- **Drivkrafter:** för vissa företag är AI redan en viktig konkurrensfaktor, medan andra ännu saknar tydliga drivkrafter och drivkrafterna är generellt sett svaga i offentlig verksamhet.
- **Datatillgång:** inom de flesta områden är bristande datatillgång en avgörande begränsning för utveckling av affärsmodeller och verksamhetsmodeller baserade på AI-tillämpningar.
- **Kompetens:** företags och offentliga verksamheters begränsade AI-kompetens, hos både ledningar och medarbetare, hämmar AI-utvecklingen i de flesta verksamheter.

Affärs- och verksamhetsmodeller, datatillgång, och kompetens är ömsesidigt beroende och påverkas starkt av varandra i företag och offentliga verksamheter. Utan tydliga perspektiv för verksamhetsnytta hämmas drivkrafterna för AI-baserade investeringar. Är verksamhetsnyttan inte tydlig uppfattas inte heller AI-kompetens som en viktig faktor för värdeskapande och effektivitet, vilket påverkar rekryteringsmönster och kompetensutveckling. Begränsad AI-kompetens, på ledningsnivå och hos

102 https://www.gartner.com/smarterwithgartner/5-trends-emerge-in-gartner-hype-cycle-for-emerging-technologies-2018
103 https://www.vinnova.se/en/publikationer/artificiell-intelligens-i-svenskt-naringsliv-och-samhalle

medarbetare, gör det i sin tur svårt att utveckla AI-baserade affärs- och verksamhetsmodeller. Datatillgång och möjligheter att kombinera olika data kommer att vara av fundamental betydelse för vilka tillämpningar som är möjliga att utveckla. Databegränsningar som försvårar eller omöjliggör utveckling av AI-baserade produkter och processer försvagar drivkrafterna för AI-investeringar.

Davenport och Dasgupta (2019)[104] delger sin syn på vad som krävs för att bygga upp en organisations AI-förmåga. Punktlistan nedan är väldigt generell och gäller för många organisatoriska förändringsarbeten, dock finns i deras text mer utvecklade exempel som kan tjäna som inspiration.

- Identify business-driven use-cases
- Determine the appropriate level of ambition
- Create a target data architecture
- Manage external innovation
- Develop and maintain a network of AI champions
- Spread success stories

AI är alltså inte en ”hemlig ingrediens” som kan läggas till vilken verksamhet som helst utan att först analysera av vad som krävs och hur verksamheten påverkas. För att lyckas med att framgångsrikt introducera AI-baserade system i en verksamhet och att få medarbetarna att förstå, så måste AI-kunskap introduceras på rätt nivå för olika grupper. Utvecklarna är intresserade av specifika AI-algoritmer och tekniken, medan verksamhetsanalytikerna behöver förstå hur affärs- och arbetsprocesserna förändras. Företags/verksamhetsledningen kanske mest intresserar sig för den ekonomiska bärkraften och även om de inte behöver vara tekniska specialister behöver de förstå olika typer av beräknings- och klassificeringsproblem. Detta för att förstå förutsättningar, ha realistiska förväntningar, och kunna fatta kloka beslut om när AI är lämpligt. Alla måste också förstå vilka krav AI-användningen ställer på tillgång till data från verksamheten.

Drift av kärnkraftverk kännetecknas av extremt höga säkerhetsnivåer och omfattande regleringar. AI-system kan få tillgång till information från långt fler källor, varav vissa som idag inte är tillgängliga eller möjliga att analysera för mänskliga operatörer. Detta kommer givetvis påverka certifierings- och ackrediteringsprocessser. Genombrotten för AI inom kärnkraften kan därför förväntas ske långsammare och kräva nya typer av reglering- och certifieringsprocesser.

Som för all förståelse och ackreditering av säkerhetskritiska digitala system behövs en detaljerad informationsmodell över systemet. Detta blir särskilt tydligt vid system som innehåller maskininlärningskomponenter och processteg där data transformeras automatiskt i flera steg på det sätt som sker i djupinlärningstillämpningar. När ett maskininlärningsbaserat system tränas upp ökar vikten av beskrivningar av hur och när systemet tränas, på vilken datamängd, och att utdata/beslut från systemet kontinerligt kontrolleras. Om detta inte är tydligt så kommer AI-tillämpningar i hög utsträckning att vara svåra att tillämpade i säkerhetskritiska verksamheter. System som lär sig över tid och hittar egna, svårgranskade beslutsgrunder lär bli mycket krävande att certifiera. De mest uppmärksammade framstegen inom AI det senaste året har utnyttjat olika former av artificiella neurala nätverk och deras förmåga att se mönster i komplexa datamängder. Detta har sedan kombinerats med moduler som använder andra AI-tekniker för att skapa agenter/beslutsstöd som användare kan interagera med.

Klustring av data, klassificering av tillstånd, och prediktioner som baserar sig på mätvärden från en anläggnings olika processer och som analyserats med någon AI-ansats, kan bli underlag för värdefulla

104 https://hbr.org/2019/01/how-to-set-up-an-ai-center-of-excellence

beslutsstöd och visualiseringar om de utformas rätt. Tillämpningar bedöms finns i flera olika typer av kontrollrum (centrala kontrollrum, bevakningscentraler och ledningscentraler) i ett kärnkraftverk. Det finns också många andra processer än den direkta styrningen av ett kärnkraftverk där AI kan vara användbart.

Datasäkerhetsfrågor finns alltid och kan bli allvarligare allteftersom AI används i större utsträckning. En god överblick över generella scenarier och hot finns via fotnot[105]. EU verksamhet för att möta dessa utmaningar beskrivs av EU via fotnot[106].

EU JRC (2018) tar upp AI ur flera perspektiv, se avsnitt 5.1, som efter fördjupad läsning alla är relevanta som embryo för eventuella framtida riktlinjer för SSM. Särskilt dataperspektivet och möjligheterna att dela med sig av data, eftersom träningsdata är en nödvändig förutsättning för flera maskininlärningsansatser. Detta kan dock naturligtvis ha implikationer för olika kärntekniska tillståndshavares kommersiella intressen.

I parafras med regeringens bedömning av AI-området (Regeringskansliet, 2018) om att Sverige behöver vara aktiva enligt punkterna nedan, gäller motsvarande men anpassat för SSM:s verksamhet, över tid:

- Sverige behöver utveckla regler, standarder, normer och etiska principer i syfte att vägleda etisk och hållbar AI och användning av AI
- Sverige behöver verka för svenska och internationella standarder och regelverk som främjar användning av AI och förebygger risker
- Sverige behöver kontinuerligt se över behovet av digital infrastruktur för att tillvarata möjligheterna som AI kan ge
- Sverige behöver fortsätta arbetet med att tillgängliggöra data som kan utgöra en samlad infrastruktur för att använda AI på områden där det tillför nytta
- Sverige behöver fortsätta att ta en aktiv roll i EU arbetet med att främja digitalisering och med att möjliggöra nyttan som användningen av AI kan medföra

AI, särskilt AI-ansatser med förmåga till maskininlärning, tenderar att vara användbara när målen är tydliga, men där vägen dit är otydlig eller för omständliga att specificera. Lyckade tillämpningar brukar karakteriseras av att antalet variabler i datamängden som AI-systemet ska hantera är relativt begränsat, så att systemet har en chans att lära sig sambandet givet tillgänglig mängd träningsdata. Chansen att lyckas gynnas också av att mängden alternativ i utdata inte är för stor. Samtidigt bör antalet kombinationer av indata och önskat utfall vara så stor att det inte är praktiskt görbart att specificera sambanden på något annat enklare sätt. Det bör också vara enkelt att kunna koppla utfallet från systemet till önskvärda och icke-önskvärda lägen så att systemet har möjligt att få kunskap om önskvärda sluttillstånd. Målen för systemets användning måste alltså vara möjliga att göra mycket explicita. Systemet kommer inte att ha en egen uppfattning om vad som är rimligt eller önskvärt om det inte ges mycket explicita måltillstånd på något sätt.

De allra svåraste frågorna kring AI-tillämpningar berör i slutändan oftast inte de tekniska utmaningarna, utan snarare filosofiska och moraliska konsekvenser utav användningen av AI. Vid all AI-introduktion tvingas utvecklare, produktägare, och användare att vara väldigt explicita vad gäller verksamhetens mål och prioriteter, vilket kan leda till svåra beslut och avdömningar som man tidigare kunnat undvika. Virginia Dignum, professor vid Umeå universitet och medlem i EU HLEG AI, använde

105 https://maliciousaireport.com
106 https://ec.europa.eu/research/sam/pdf/sam_cybersecurity_report.pdf

följande ord vid Vetenskapsrådets seminarie kring ”Ansvarsfull teknisk utveckling – AI, robotik, etik”, den 8 mars 2019:

- AI påverkar och är påverkad av våra sociala system
- Design är aldrig värdeneutral
- Öppenhet och tydlighet är centralt – ansvarsskyldighet *(accountability)*, ansvarsfullhet *(responsibility)*, och transparens *(transparency)* är ledord för utvecklingen inom området
- Optimal AI är förklaringsbar AI
- AI-system är artefakter som utvecklas av oss för våra egna syften
- Vi sätter begränsningarna

I princip handlar det alltså inte om vad AI-system kan göra, utan om vad vi vill att de ska göra.

Vid tidpunkten för denna rapport så har inga offentligt dokumenterade exempel på implementation av AI inom operativ kärnkraftssäkerhet kunnat identifieras. Med anledning av detta så kan man i nuläget endast spekulera i hur AI teknik i framtiden skulle kunna tänkas komma att implementeras för detta syfte. Utifrån tillämpningar i andra, jämförbara domäner, så är en rimlig spekulation att det inom operativ kärnkraftssäkerhet är troligt att de första tillämpningarna kommer ske inom ett eller flera av nedanstående exempel:

- Processövervakning/-optimering (*big data analysis*)
- Avvikelsedetektion/-hantering
- Prediktiv underhållsplanering
- Planeringsverktyg för icke säkerhetskritiska arbetsprocesser
- Digital assistent (fråga/svar system)

8 Referenser

AlgoAware (2018). AlgoAware: State-of-the-Art Report - Algorithmic decision-making. https://www.algoaware.eu/wp-content/uploads/2018/08/AlgoAware-State-of-the-Art-Report.pdf

Armstrong, S., Sotala, K., & ÓhÉigeartaigh, S.S. (2014). The errors, insights and lessons of famous AI predictions – and what they mean for the future. Journal of Experimental & Theoretical Artificial Intelligence. http://www.fhi.ox.ac.uk/wp-content/uploads/FAIC.pdf

Bradshaw, J.M., Hoffman, R., Johnson, M., & Woods, D. (2013). Seven deadly myths of autonomous systems. IEEE Intelligent Systems, 28(3), 54–61.

Chollet, F. (2017). Deep Learning with Python. Manning. https://www.manning.com/books/deep-learning-with-python

Chen, F. (2016). AI, Deep Learning, and Machine Learning: A Primer. http://a16z.com/2016/06/10/ai-deep-learning-machines

Darling, M. C, Luger, G.F., Jones, T.B., Denman, M.R. & Groth, K.M. (2018). Intelligent Modeling for Nuclear Power Plant Accident Management. International Journal on Artificial Intelligence Tools. Vol. 27, No. 02, 1850003. https://www.worldscientific.com/doi/abs/10.1142/S0218213018500033

DoD (2012). The Role of Autonomy in DoD Systems. US Department of Defense. https://fas.org/irp/agency/dod/dsb/autonomy.pdf

Davenport, T.H., & Dasgupta, S. (2019). How to Set Up an AI Center of Excellence. https://hbr.org/2019/01/how-to-set-up-an-ai-center-of-excellence

Domingos, P. (2015). The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. New York: Basic Books

Elsevier (2018). Artificial Intelligence: How knowledge is created, transferred, and used. https://www.elsevier.com/__data/assets/pdf_file/0010/823654/ACAD-RL-AS-RE-ai-report-WEB.pdf

EU (2017). Digital Transformation Monitor AI Policy Seminar: Towards an EU strategic plan for AI. 29 November 2017. https://ec.europa.eu/growth/tools-databases/dem/monitor/sites/default/files/Main%20findings%20of%20the%20Policy%20Seminar%20FINAL2_0.pdf

EU HLEG AI (2018 a). Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe, Brussels, 25.4.2018 COM(2018) 237 final. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56018

EU HLEG AI (2018 b). The European Commission's high-level expert group on Artificial Intelligence. A definition of AI – Main capabilities and scientific disciplines. https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december.pdf

EU HLEG AI (2019). The European Commission's high-level expert group on Artificial Intelligence. Ethics guidelines for trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

EU JRC (2018). Artificial intelligence – European Perspective. JRC113826 EUR 29425. DOI 10.2760/11251. https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/artificial-intelligence-european-perspective

Floridi, L., Cowls, J., Beltrametti, M. m.fl. (2018). Minds & Machines, 28: 689. https://link.springer.com/article/10.1007%2Fs11023-018-9482-5

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. http://www.deeplearningbook.org

Goodfellow, I., Pouget-Abadie, P., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. https://arxiv.org/pdf/1406.2661.pdf

Hollnagel, E., & Woods, D. (2005). Joint Cognitive Systems -Foundations of Cognitive Systems Engineering. Boca Raton, FL: Taylor & Francis.

Jastrow, R. (1982). The thinking computer. Science Digest, 90:6, 54–55. 106-107.

Johnson, M. (2014). Coactive Design: Designing Support for Interdependence in Human-Robot Teamwork. PhD thesis Technische Universiteit Delft.

Karras, T., Laine, S., & Aila, T. (2018). A Style-Based Generator Architecture for Generative Adversarial Networks. https://arxiv.org/pdf/1812.04948.pdf

Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten challenges for making automation a "team player" in joint human-agent activity. IEEE Intelligent Systems, 19(6), 91–95.

Lackman, T. (2011). Utredning och kartläggning av tillfällen då människan räddat och förbättrat en situation där automationen inte räckt till eller fungerat fel. SSM rapport 2011:24. https://www.stralsakerhetsmyndigheten.se/contentassets/26e636ad39a14a90a9022b8e53f75285/201124-utredning-och-kartlaggning-av-tillfallen-da-manniskan-raddat-och-forbattrat-en-situation-dar-automatiken-inte-rackt-till-eller-fungerat-fel

Lin, L.J. (1993). Reinforcement learning for robots using neural networks. PhD Thesis Carnegie Mellon University. https://pdfs.semanticscholar.org/54c4/cf3a8168c1b70f91cf78a3dc98b671935492.pdf

Moedas C. (2017). Media in the Age of Artificial Intelligence. Speech at the STOA Annual Lecture, 21 November. https://ec.europa.eu/commission/commissioners/2014-2019/moedas/announcements/stoa-annual-lecture-media-age-artificial-intelligence_en

Mueller, S.T., Hoffman, R.T., Klein, G., Miller, T., & Aha, D. (2018). Panel on Explainabie Artificial Intelligence (XAI). Proceedings of the 23rd International Command and Control Research & Technology Symposium (ICCRTS), 6–9 Nov, Pensacola, FL, USA. https://www.researchgate.net/publication/329352442_Mueller_S_T_2018_Panel_on_Explainabie_Artificial_Intelligence_XAI_in_Proceedings_of_the_23rd_International_Command_and_Control_Research_Technology_Symposium_ICCRTS_6-9_Nov_Pensacola_FL_USA

Mullins, C. (2012). Retrospective Analysis of Technology Forecasting. https://apps.dtic.mil/dtic/tr/fulltext/u2/a568107.pdf

Regeringskansliet (2018). Nationell inriktning för artificiell intelligens. https://www.regeringen.se/49a828/contentassets/844d30fb0d594d1b9d96e2f5d57ed14b/2018ai_webb.pdf

Russel, S., & Norvig, P. (2010). Artificial Intelligence: A Modern Approach 3rd Ed. http://aima.cs.berkeley.edu

Searle, J. (1980). Minds, Brains, and Programs. Behavioral and Brain Sciences. http://cogprints.org/7150/1/10.1.1.83.5248.pdf

Schubert, K. (2017). Artificiell Intelligens för Militärt Beslutsstöd. FOI rapport FOI-R--4552--SE. https://www.foi.se/rapportsammanfattning?reportNo=FOI-R--4552--SE

Sutton, R.S., & Barto, A.G. (2018). Reinforcement Learning: An Introduction 2nd Ed. The MIT Press Cambridge, Massachusetts. http://incompleteideas.net/sutton/book/RLbook2018trimmed.pdf

Uhrig, R. (1989). The use of artificial intelligence to enhance the safety in nuclear power plants. Proceedings of International Nuclear Information System (INIS).

Vinnova (2018). Artificiell intelligens i svenskt näringsliv och samhälle. Delrapport 2018-02-12, dnr 2017–05616. https://www.vinnova.se/en/publikationer/artificiell-intelligens-i-svenskt-naringsliv-och-samhalle

2019:29

Strålsäkerhetsmyndigheten har ett samlat ansvar för att samhället är strålsäkert. Vi arbetar för att uppnå strålsäkerhet inom en rad områden: kärnkraft, sjukvård samt kommersiella produkter och tjänster. Dessutom arbetar vi med skydd mot naturlig strålning och för att höja strålsäkerheten internationellt.

Myndigheten verkar pådrivande och förebyggande för att skydda människor och miljö från oönskade effekter av strålning, nu och i framtiden. Vi ger ut föreskrifter och kontrollerar genom tillsyn att de efterlevs, vi stödjer forskning, utbildar, informerar och ger råd. Verksamheter med strålning kräver i många fall tillstånd från myndigheten. Vi har krisberedskap dygnet runt för att kunna begränsa effekterna av olyckor med strålning och av avsiktlig spridning av radioaktiva ämnen. Vi deltar i internationella samarbeten för att öka strålsäkerheten och finansierar projekt som syftar till att höja strålsäkerheten i vissa östeuropeiska länder.

Strålsäkerhetsmyndigheten sorterar under Miljödepartementet. Hos oss arbetar drygt 300 personer med kompetens inom teknik, naturvetenskap, beteendevetenskap, juridik, ekonomi och kommunikation. Myndigheten är certifierad inom kvalitet, miljö och arbetsmiljö.

Strålsäkerhetsmyndigheten
Swedish Radiation Safety Authority

SE-171 16 Stockholm
Tel: +46 8 799 40 00
E-mail: registrator@ssm.se
Web: stralsakerhetsmyndigheten.se