Avskanning av dataanalys och AI

Rapport år 2023

RONNIE JOHANSSON (RED.), PETER HAMMAR,
ALEXANDER KARLSSON, SIDNEY RYDSTRÖM

FOI-R--5483--SE
ISSN 1650-1942 Juli 2023

Ronnie Johansson (red.), Peter Hammar,
Alexander Karlsson, Sidney Rydström

Avskanning av dataanalys och AI

Rapport år 2023

Titel	Avskanning av dataanalys och AI – Rapport år 2023
Title	Horizon Scanning of Data Analysis and AI – Report in 2023
Rapportnr/Report no	FOI-R--5483--SE
Månad/Month	Juli
Utgivningsår/Year	2023
Antal sidor/Pages	60
ISSN	1650-1942
Uppdragsgivare/Client	FMV
Forskningsområde	Ledningsteknologi
FoT-område	Inget FoT-område
Projektnr/Project no	E86266
Godkänd av/Approved by	Linda Sjödin
Ansvarig avdelning	Cyberförsvar och ledningsteknik

Bild/Cover: Shutterstock

Sammanfattning

Dataanalys och det närbesläktade området artificiell intelligens utvecklas snabbt och är av stort intresse för ledningsområdet då de kan erbjuda åtråvärda egenskaper som automatiserad hantering av stora datamängder och modeller för att tolka ny data. Den här rapporten omfattar en avskanning av fyra utvalda delområden: preskriptiv analys, osäkerhetshantering, förklarbarhet (XAI) och systemperspektiv. Rapporten innehåller dessutom en uppskattning av den framtida utvecklingen av delområdena och en jämförelse med avseende på deras mognadsgrad och relevans.

Nyckelord: artificiell intelligens (AI), dataanalys, avskanning, preskriptiv dataanalys, osäkerhetshantering, förklarbarhet (XAI)

Summary

Data analysis and the closely related field of artificial intelligence are developing rapidly and are of great interest to the command and control field as they can offer desirable features such as automated handling of large data sets and models for interpreting new data. This report includes a scan of four selected sub-areas: prescriptive analysis, uncertainty management, explainability (XAI) and systems perspective. The report also contains an estimate of the future development of the sub-areas and a comparison with respect to their degree of maturity and relevance.

Keywords: artificial intelligence (AI), data analysis, horizon scanning, prescriptive data analysis, uncertainty management, explainable AI (XAI)

Innehållsförteckning

1 Inledning

Rapporten omfattar en avskanning av aktuell forskningslitteratur inom dataanalys och artificiell intelligens med fokus på utvalda delområden.

1.1 Syfte

Syftet med rapporten är att skanna av forskningsområdena *dataanalys* (eng. *data analysis*) och *artificiell intelligens* (eng. *artificial intelligence*) vilka är till nytta för ledningsområdet. Rapporten omfattar även framtidsprognoser och analys för respektive område med avseende på försvarstillämpningar.

1.2 Metod

Avskanningen av områdena utfördes parallellt av författarna, men angreppssätten har varierat. Någon generell allomfattande avskanning har inte skett utan författarna har valt delområden att studera baserat på deras individuella expertis. De utvalda delområdena är *preskriptiv analys*, *osäkerhetshantering*, *förklarbarhet* och *systemperspektiv*.

Generellt sett har meta-sökmotorn Web of Science[1] (WoS) nyttjats. Urvalet har bland annat omfattat ämnesspecifika tidskrifter och översiktsartiklar. En WoS-sökfråga (som gäller titlar och sammanfattningar, TS) kan exempelvis se ut så här:

TS = (("X" OR "Y NEAR/2 Z") AND "Q")

där X avser ett nyckelord, exempelvis det aktuella området, och Y NEAR\2 Z ett alternativt nyckelord Y på två ordavstånd från ett följdnyckelord Z. Slutligen måste texten även innehålla nyckelordet Q. Efter urval av artiklar, inläsning och sammanfattning följde en uppskattning av områdenas tillstånd och framtidsprognos. Även Google Scholar[2] har använts.

1.3 Bakgrund

Dataanalys och artificiell intelligens (DA and AI) är två datalogiska områden som är både omfattande och överlappande. Ett antal olika definitioner av dessa områden finns, men det perspektiv som vi antar i den här rapporten förklaras nedan.

DA avser vanligtvis processen att undersöka och tolka data för att extrahera insikter eller dra slutsatser från dem. Detta kan innebära förberedande uppgifter som datarensning (formatera och fixa till fel och brus i data så att den är lämplig för algoritmisk analys); analyssteg som datavisualisering och statistisk analys; och efterbearbetningssteg som att skapa rapporter och lagra data för senare användning.

DA syftar till att organisationer skall kunna ta snabbare, bättre och mer välinformerade beslut med målet att skapa värde för organisationen. I den engelskspråkiga litteraturen används ibland även ordet *data analytics* för att omfatta förarbetnings-, analys- och efterarbetningsstegen. I de sammanhangen avser data analysis enbart analyssteget.

DA delas in i tre[3] underkategorier: *deskriptiv* (eng. *descriptive*), *prediktiv* (eng. *predictive*) och *preskriptiv* (eng. *prescriptive*) *dataanalys*. Deskriptiv dataanalys svarar på frågan ”Vad har hänt?”. Typiskt används tekniker så som dataaggregering/-klustring och data mining. Prediktiv dataanalys svarar på frågan ”Vad kommer att hända?” Typiskt används

[1] https://www/webofscience.com/ (besökt juni 2023)

[2] https://scholar.google.com/ (besökt juni 2023)

[3] Ibland används en fjärde kategori, diagnostisk analys.

maskininlärning som producerar prediktioner och prognoser. Preskriptiv dataanalys svarar på frågan ”Hur får vi det att hända?” eller ”Vad kan vi göra?”

Data analytics benämns i forskningslitteraturen ofta *business analytics*, eller *analytics* med eller utan benämning av tillämpningsområdet[4]. Det är ett tämligen moget forskningsområde, och huvuddelen av välciterade forskningsrapporter handlar om användningen av en eller en kombination av flera existerande metoder på olika tillämpningsfall.

DA omfattar alltså metoder som är baserade på data. AI kan ses som ett bredare område som i stora drag omfattar dataanalys. AI-metoder behöver exempelvis inte vara baserade på tillgång till en viss datamängd utan kan istället dra nytta av formaliserad (datoranpassad) expertkunskap.

AI hänvisar till maskiners förmåga att utföra uppgifter som normalt kräver mänsklig intelligens – till exempel att känna igen mönster, lära av erfarenhet, dra slutsatser, göra förutsägelser eller vidta åtgärder – oavsett om det är digitalt eller som den smarta mjukvaran bakom autonoma fysiska system (”Artificial Intelligence – Air Force Research Laboratory” 2023).

Notera att DA och AI har olika syften. DA är fokuserat på att extrahera värdefull information (exempelvis för att stödja mänskligt beslutsfattande) från en datamängd, medan AI kan omfatta skapandet av intelligent beteende (exempelvis för robotar och agenter av olika slag).

1.4 Läsanvisningar

I kapitel 2-5 redovisas avskanningarna av de fyra utvalda delområdena, ett för varje kapitel. Kapitlen har en liknande struktur med en inledande del med grundläggande information om delområdet i fråga (eventuellt med ytterligare en nedbrytning av delområdet i underområden), en genomgång av intressanta forskningsartiklar, samt en avslutande framtidsprognos och analys av områdets nytta för försvarstillämpningar.

I kapitel 2 diskuteras preskriptiv analys som är den del av DA som resonerar kring möjliga beslut och handlingsalternativ. I kapitel 3 diskuteras osäkerhetshantering med fokus på maskininlärning. I kapitel 4 diskuteras förklarbarhet, dvs angreppssätt för att göra AI-metoder mer begripliga och därmed trovärdiga för en beslutsfattare. I kapitel 5 diskuteras AI-metoder ur ett systemperspektiv och hur dessa skiljer sig från annan mjukvara och vilka utmaningar som det medför. Avslutningsvis, i kapitel 6, sammanfattas och värderas de områden som diskuterats i kapitel 2-5 med avseende på mognadsgrad och relevans för försvarstillämpningar.

Rapportens språk är svenska och i största möjliga mån har viktiga etablerade engelskspråkiga begrepp översatts till svenska då detta ger ett mer enhetligt och flytande språk. I många fall finns inte någon vedertagen svensk översättning, men i dessa fall introduceras ändå en översättning för användning i rapporten. Den ursprungliga engelskspråkiga termen finns alltid med för spårbarhetens skull.

[4] T.ex. Management, Operational, Crime, People, Healthcare, Web Analytics

2 Preskriptiv analys

I rapportens inledning (avsnitt 1.3) delas dataanalysområdet upp i delområdena deskriptiv, prediktiv och preskriptiv dataanalys. I det här kapitlet utforskas några typer av metoder för det sistnämnda. I preskriptiv dataanalys (PDA) används olika tekniker för att utvärdera och hitta bästa alternativ för en beslutsprocess givet en mängd mål, krav och begränsningar (Frazzetto m.fl. 2019).

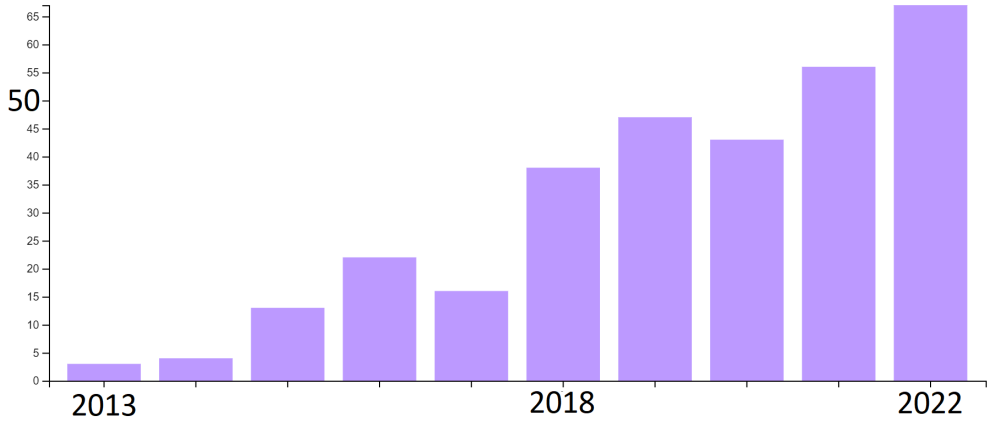
Fram till nyligen har fokus i akademin och industrin varit på deskriptiv och prediktiv dataanalys. PDA, vilket syftar till att ta fram de bästa handlingsalternativen eller beslutsalternativen, är nästa steg mot mognad inom dataanalys. PDA drar typiskt nytta av resultat från deskriptiv och prediktiv dataanalys, men behöver även en uppfattning om hur lämpliga beslutsalternativ formuleras och värderas. Det avser leda till optimala beslut med framförhållning för att öka nyttan. Annorlunda uttryckt, PDA syftar till att identifiera de bästa handlingsalternativen för optimalt beslutsfattande beträffande organisationens framtida resultat.

PDA svarar på frågorna *vad skall göras?* och *varför?* Området är besläktat med operationsanalys, och rör sig nu från tillämpningsfall med matematisk optimering och simulering till alltmer användande av maskininlärningsmetoder. Preskriptiv innebär att föreslå det bästa beslutsalternativet utifrån den förutspådda framtida utvecklingen, baserat på stora datamängder (Lepenioti m.fl. 2021). Begreppet ”beslut” är centralt för PDA. Även i prediktiv analys tas beslut, dvs. beslut om exempelvis klass vid klassificering av indata. Viktigt är därför att notera att det för PDA handlar om beslut om handlingsalternativ, vilket tillför PDA:s karaktäristiska dimension, dvs vilka handlingar som är möjliga och hur de skall värderas.

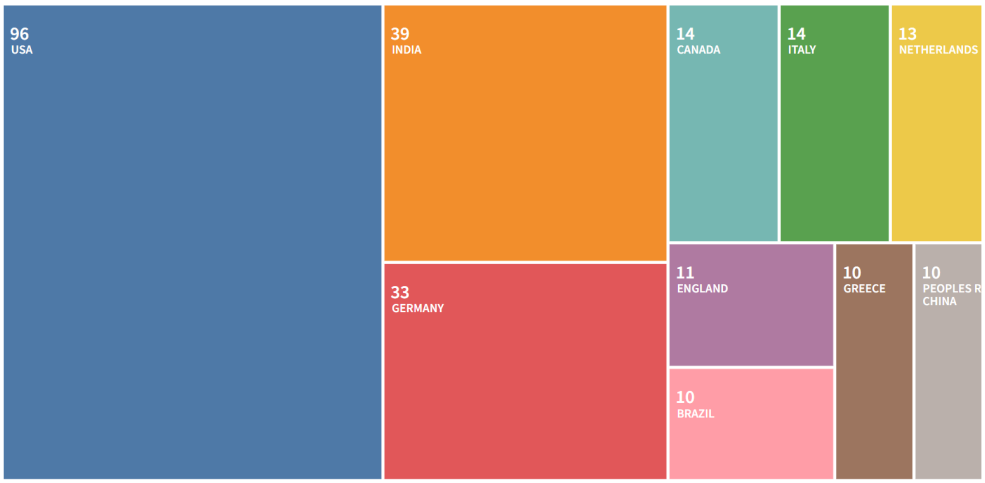
PDA har de senaste åren erhållit alltmer fokus (Lepenioti m.fl. 2020) även om det är från en förhållandevis låg nivå. I figur 1 visas antal publikationer om ”prescriptive analytics” som hittades (i titel, sammanfattning och nyckelord) med hjälp av WoS[5] och i figur 2 forskningens ursprungsländer. Lepenioti (2020) påpekar att PDA inkorporerar resultatet av den prediktiva dataanalysen och drar dessutom nytta av AI-metoder, optimeringsalgoritmer och expertsystem. Målet är att ta fram beslut som är anpassningsbara, automatiserade, begränsade, tidsberoende och optimala. Lepenioti m.fl. (2020) klassificerar metoder som används inom PDA i ett antal delområden:

- Sannolikhetsmodeller
- Maskininlärning och data mining
- Matematisk programmering
- Evolutionära beräkningar
- Simulering
- Logik-baserade metoder

[5] WoS-sökfråga: (TS=(prescriptive analytics) AND (SU = (Computer Science) OR SU = (Mathematics) OR SU = (Automation & Control Systems) OR SU = (Engineering) OR SU = (Remote Sensing) OR SU = (Robotics) OR SU = (Science & Technology Other Topics) OR SU = (Telecommunications))) AND PY=(2012-2022)



Figur 1. Antal forskningsartiklar med nyckelordet ”prescriptive analytics” under perioden 2013-2022



Figur 2. Forskningens ursprungsländer för ”prescriptive analytics” 2013–2022

Viktigt att notera är också att AI erbjuder metoder som inte bara tar fram ”engångsbeslut”, utan även mer långsiktiga som policyer (dvs. modeller som föreslår beslut för olika tänkbara tillstånd) (Sutton och Barto 2018) och planer (desJardins m.fl. 1999).

I resten av kapitlet undersöks hur forskningen inom två metodtyper, multikriterieanalys (avsnitt 2.1) och POMDP:er (avsnitt 2.2), kan bidra till PDA.

2.1 Multikriterieanalys

Preskriptiv dataanalys går hela vägen från data till bästa handlingsalternativet eller beslutet i en given situation. Det kan således ses som en typ av datadrivet beslutsfattande, men det bör understrykas att många metoder inom området inte kräver data, utan kan i stället baseras på expertkunskap. Praktiska beslutssituationer innehåller ofta komplexa avvägningar mellan olika alternativ som inte är enkla att jämförbara, och vilar på många olika bedömningsgrunder. Multikriterieanalys (eng. multi-criteria decision making, MCDM) är samlingsnamn för metoder som hanterar just sådana beslutssituationer. MCDM är ett relativt moget forskningsfält med gamla anor men det finns en starkt ökande trend i antalet publikationer under 2000-talet och framåt. MCDM lyfts här därför att användandet av dessa metoder inom dataanalys är relativt nytt. De MCDM-metoder som används är i huvudsak väletablerade. De används både integrerat med dataanalys-baserade metoder så som data mining, maskininlärning eller text mining-tekniker, eller fristående utifrån framtagna datamängder (Yalcin, Kilic, och Delen 2022).

I MCDM jämförs och väljs det bästa alternativet från en mängd beslutsalternativ som involverar flera kriterier eller mål, vilka ofta i sig inte är direkt jämförbara. Fältet delas in i

beslut som grundas på flera egenskaper, multi-attribute decision making (MADM), där beslutsalternativen är uppräkneliga och ofta ett begränsat antal, och beslut som fattas givet flera olika målsättningar, multi-objective decision making (MODM), där beslutsalternativen kan var oändligt många. MADM-metoder resulterar normalt i en lösning eller rankning av flera lösningar som grundas i beslutssituationens egenskaper. I MODM är optimeringen över kontinuerliga variabler och går ut på att hitta den bästa lösningen som optimerar samtliga målsättningar. De ger ofta en mängd pareto-optimala lösningar som svar, som avvägning mellan de olika målen.

Det finns en mängd olika metoder inom MADM, och vilken som lämpar sig bäst i en viss situation beror på tillämpningen. De MADM-metoder som används inom dataanalys kan delas in i sex klasser, parvis jämförande, överträffande, avståndsbaserade, interaktionsbaserade, nyttobaserade och övriga metoder. Inom preskriptiv analys är avståndsbaserade metoder mycket använda, och specifikt TOPSIS-metoden (Hwang och Yoon 1981). MODM-metoder inom dataanalys kan delas in i fyra kategorier utifrån beslutsfattarens roll, metoder utifrån avsaknad av beslutsfattarens preferens, preferens a priori, preferens a posteriori samt interaktiva metoder. (Yalcin, Kilic, och Delen 2022)

Yalcin (2022) har gjort en gedigen genomgång av publikationer kring multikriterieanalys kopplat till dataanalys. De flesta artiklarna utgör tillämpningar av etablerade metoder på nya problem, antingen med en eller en kombination av flera algoritmer. Det är alltså få verk som presenterar någon nyutvecklad algoritm. Därtill fokuserar en mängd av artiklarna på specifik optimering kring ett problem, och bedöms inte vara enkelt generaliserbart till en konkret (militär) beslutssituation. Yalcin och medförfattare noterar att MCDM används både som input till deskriptiv och prediktiv dataanalys samt för preskriptiv analys när det specifikt används som ett medel för beslutsfattande för att lista och välja alternativ. MCDM-metoden kan användas för optimering generellt, men i avseendet preskriptiv analys används MCDM specifikt för val av handlingsalternativ i en beslutssituation. Utifrån detta har ett par artiklar valts ut, utifrån både Yalcins genomgång och en litteratursökning[6], som lyfter nya algoritm- och metodutvecklingar inom området, och vi låter dem här nedan exemplifiera forskningsläget inom multikriterieanalys tillämpbar inom preskriptiv dataanalys.

2.1.1 Bipolär modell för förbättringsplanering

Shen och Tzeng (Shen och Tzeng 2016) har utvecklat en ny modell för multikriterieanalys. Studien presenterar en ny MADM-modell som hanterar hela kedjan från rankning och urval av alternativ till förbättringsplanering, baserat på alternativens likhet med positiva kontexter och olikhet från negativa sådana. Kontexter är i detta sammanhang alla de förhållanden och tillstånd som råder i en beslutssituation och påverkar beslutsfattandet. Beslutsregler och MADM-metoder kombineras vilket ger både en rankning eller val av beslutsalternativ, och skapar ett lättförståeligt underlag för förbättringsplanering.

Metoden illustreras med ett exempel där historisk ekonomisk data används för att ta fram beslutsalternativ och vägleda inom vilka områden ett företag skall fokusera för att förbättra sig gentemot beslutskriterierna. Den föreslagna modellen kan hjälpa ett företag att utveckla analys till prioriterade kontexter, som kan användas för att vägleda förbättringsplanering. Modellen används föra att lösa utmaningen med att förutse finansiell prestanda och erhålla vägledning för att förbättra lönsamhet.

Varje kontext i modellen beskriver ett scenario där företaget kan utvecklas (genom att sträva efter att vara mer lik de positiva eller mindre lik de negativa kontexterna). Prioriterade kontexter erhålles vilka kan hjälpa företaget att välja de mål (kontexter) som bidrar mest för utvecklingen givet begränsade resurser. Detta lyfts fram som att kunna anses vara en integration mellan MADM och scenarioplanering.

[6] Högst citerade artiklar indexerade i Web of Science, för topic = ”MCDM” eller ”multi criteria decision making” från 2018 och framåt.

Studien presenterar en bipolär beslutsmodell, som använder DRSA (dominance-based rough-set approach) och en beslutsmetod som avväger motstridiga kriterier (modified VIKOR). Bipolär innebär i sammanhanget att beslutsregler indelas i två grupper: en positiv (för alternativ som är fördelaktiga eller accepterade) och en negativ (för ofördelaktiga eller oönskade alternativ). Beslutsalternativ rankas utifrån deras likhet med den positiva gruppen och dess olikhet med den negativa.

Studien försöker transformera DRSA-beslutsregler till en integrerad fuzzy MADM-modell. Den föreslagna modellen förväntas öka förmågan hos DRSA till att ranka eller välja (även inom en specifik beslutsgrupp eller odefinierad beslutsgrupp), och utökar därmed tillämpningarna av DRSA till beslutsvetenskap och dataanalys. Metoden kan alltså ranka alternativ som ligger i samma beslutsklass.

Modellen består i huvudsak av tre delar: DRSA, bipolär modell och modifierad VIKOR. DRSA kan dela in objekt i en datamängd i olika beslutsklasser baserat på dess attribut. Dominansförhållanden mellan objekten arbetas fram, vilka sedan används för att ta fram beslutsregler. DRSA tar också fram de attribut som är av störst betydelse.

Beslutsreglerna (kontexterna) från DRSA transformeras till en bipolär modell med hjälp av viktning. En stödjande vikt är ett mått på hur ofta en specifik kontext förekommer, med andra ord, ju fler datapunkter som uppfyller en viss beslutsregel desto högre vikt. Ju högre vikt desto större förtroende bör en beslutsfattare ha för den givna regeln, baserat på observerade data.

Oskarpa utvärderingar (baserat på oskarpa mängder, eng. fuzzy sets) används sedan för att väga in kontexterna från den bipolära modellen in i en modifierad VIKOR-metod. Den modifierade VIKOR-metoden (Opricovic och Tzeng 2007) är en multikriterieanalysmetod som används för att ranka alternativ utifrån hur nära de är den ideala lösningen. Den tar hänsyn till flera kriterier och väger samman dessa på ett balanserat sätt, beaktande både enskilda och samlade mängden kriterier.

Den bipolära modellen med den modifierade VIKOR-metoden resulterar inte bara i en rankning av beslutsalternativ, men ger också stöd för förbättringsplanering. Det är denna djupgående analys som skapar underlag för förbättringar som är det huvudsakliga bidraget i artikeln. Motsvarande information uppges inte vara möjlig att erhålla med konventionella statistiska metoder.

Den föreslagna metoden utökar existerande metoder (DRSA) med att kunna ranka alternativ inom samma beslutsklass. Därtill erhålles en kontextuell vinkling för systematiska förbättringar. Med den modifierade VIKOR-metoden kan den föreslagna modellen stödja ett företag i att transformera analys till prioriterade kontexter, vilka kan vägleda förbättringsplanering. De i artikeln erhållna resultaten sägs överbrygga tillämpningar av datadriven analys till beslutsvetenskap i praktiken.

2.1.2 MCDM och klusteranalys

I en studie av Maghsoodi och medförfattare (Ijadi Maghsoodi m.fl. 2018) presenteras en metod för att förbättra MCDM baserat på klusteranalys av en stor datamängd med skarpa data. Metoden benämns CLUS-MCDA. Den går ut på att kombinera k-means klustring av data med MULTIMOORA-tekniken inom MADM. Genom metoden reduceras en stor mängd beslutsval till några slutgiltiga kandidater. Bidraget är att ta fram en MADM-metod som kan hantera stora datamängder. Artikeln visar ett tillämpningsfall som går ut på att välja ut leverantörer utifrån en stor datamängd med beskrivning av dessa (inom områden som valts av experter).

Klustring är ett sätt att automatiskt gruppera stora mängder data i ett fåtal grupper där individer i samma grupp liknar varandra. MOORA (Multi-objective optimization by ratio analysis) är en MCDM-metod som bygger på förhållandet mellan varje alternativ till det bästa eller sämsta alternativet. Detta beräknas som en poäng för varje kriterium för sig och den bästa lösningen är den som har högst sammanlagd poäng. MULTIMOORA (multiplicative

MOORA) är en utvidgning av denna metod som också tar hänsyn till hur viktiga olika kriterier är i förhållande till varandra.

I CLUS-MCDA klustras data och sedan används MULTIMOORA-metoden för att ranka alternativen i respektive kluster. Därmed kan grupper av data med liknande beslutskriterier identifieras, det blir alltså enklare att identifiera och jämföra beslutsalternativen, och det bästa alternativet i respektive kluster kan identifieras.

2.1.3 Avvägningslösning för MCDM-problem

Yazdani och medförfattare (Yazdani m.fl. 2019) föreslår en ny MCDM-metod, som benämns *Combined compromise solution*, CoCoSo. Algoritmen bidrar med en process för vikt-aggregering i *grey relational generation approach* (GRA). GRA är utvecklat för att hantera ofullständig, osäker eller motstridig information, och används för att bedöma likheten mellan datamängder. Varje alternativ jämförs med den ideala lösningen. Detta liknar VIKOR-metoden, men med en något förändrad formel. Däremot används en annan aggregeringsfunktion, med exponenten av vikterna, vilket leder till ett starkare avståndsmått. Metoden använder en jämförelsesekvens, varefter vikterna aggregeras på två sätt. Den ena använder den vanliga multiplikationsregeln, medan den andra utnyttjar viktade exponenter av avståndet från jämförelsesekvensen. Inom MCDM är en jämförelsesekvens en ordnad lista av beslutsalternativen, där ordningen i listan bestäms genom att gå igenom kriterierna och uppdatera listan så att ett alternativ hamnar före ett annat om det har en bättre utvärdering av det aktuella kriteriet.

En nyhet i metoden är att den validerar rankningen med tre olika mått, vilket skapar en aggregeringsstrategi. En kumulativ ekvation används i slutändan för att välja en rankning. Yazdani m.fl. (Yazdani m.fl. 2019) hävdar att det inte finns någon annan MCDM-metod som använder sig av denna typ av aggregering. CoCoSo använder mått från WSM (Weighted Sum Model) och WPM (Weighted Product Model) på tre sätt, aritmetiska summan, summan relativt den bästa, samt en balanserad (med en parameter som väljs av beslutsfattarna) summa av poängen från WSM och WPM. WSM bygger på att värdet för varje kriterium multipliceras med en vikt relativ till kriteriets betydelse. Den slutgiltiga poängen ges av en multiplikativ faktor och en additiv faktor, vardera av de tre måtten.

2.1.4 Ny dynamisk MCDM-metod

I traditionell MCDM antas kriterier och beslutsalternativ vara fixa, men i många fall måste föränderliga kriterier och alternativ kunna hanteras, vilket kallas för dynamisk MCDM. Artikeln av Thong m.fl. (Thong m.fl. 2020) utgör ett exempel på en metod för att hantera fallet där både kriterier, alternativ och beslutsfattare förändras över tiden för beslutsprocessen.

Bidraget använder sig av neutrosofiska mängder (eng. neutrosophic sets), vilket är en utökning av oskarpa mängder (eng. fuzzy sets). De hanterar uppdelning där ett element kan vara del i mängden, inte del i mängden, eller obestämbar. Detta används för att hantera osäkerhet och vaghet i beslutsfattande. Förändringar över tid kan hanteras med intervall av neutrosofiska mängder, och MCDM för dynamiska intervall-värden på sådana mängder (DIVNS) har tidigare tagits fram (Thong m.fl. 2019). Thong m.fl. (2019) tog fram DIVNS och de operatorer som krävs för denna mängd, samt utvecklade TOPSIS-metoden för DIVNS. I den artikeln har dock inte problemet med föränderliga kriterier, alternativ och beslutsfattare lösts, vilket görs här.

TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) är en av de mest använda metoderna för MCDM. Den baseras på att alternativ rankas utifrån deras närhet till den ideala lösningen och största avstånd till den negativa ideala lösningen. Kriterier vägs utifrån hur viktiga de är, och bästa och sämsta val för respektive kriterium tas fram varefter en rankning kan göras.

Artikeln presenterar en anpassning av TOPSIS-metoden för att hantera dynamisk MCDM med DIVNS. Artikeln exemplifierar sin metod med ett typfall där universitetsstudenter rankas. Författarna definierar ett ”generalized dynamic interval-valued neutrosophic set (GDIVNS)”, och några operatorer för avstånd och viktad aggregering för denna mängd. Utifrån dessa operatorer utvecklas ett ramverk för dynamisk TOPSIS i GDIVNS-miljön. DIVNS utvidgas med ”hesitant fuzzy sets” för att hantera fallet där kriterier, alternativ och beslutsfattare förändras över tid.

2.2 POMDP

I preskriptiv dataanalys önskar man hitta beslut och handlingar som optimerar något framtida värde, så som ett företags vinst eller utgången av en väpnad konflikt. Många beslutssituationer är återkommande, en stab måste periodiskt fatta beslut om hur förbandets resurser skall fördelas, och en robot måste kontinuerligt navigera utifrån en föränderlig omvärld. I de flesta praktiska situationer finns det också yttre faktorer vilka beslutsfattaren inte kontrollerar, men som påverkar resultatet av en handling. En ytterligare faktor som påverkar beslutsfattandet är om fullständig information om situationen inte finns att tillgå. Ett sätt att hantera dessa utmaningar är genom att modellera beslutsproblemet med hjälp av markovska beslutsprocesser (Markov Decision Process, MDP) och specifikt partiellt observerbara MDP (Partially Observable MDP, POMDP).

MDP och POMDP introducerades på 50- och 60-talen men är fortfarande aktuella, vilket tyder på dess kraft och tillämpbarhet (Chadès m.fl. 2021). I det följande görs en kort introduktion till MDP och POMDP, varefter några nya forskningsrön inom området presenteras.

2.2.1 Markov decision process

Sekventiella beslutssituationer kan hanteras med hjälp av Markov Decision Process (MDP), en matematisk metod som kombinerar markovkedjor och beslutsteori. Det är användbart i fall där resultatet av beslutet är osäkert. Beslutsfattandet sker i distinkta tidssteg, utifrån det tillstånd som råder fattas ett beslut om en handling. Handlingen påverkar övergången till ett nytt tillstånd, och resulterar i en belöning eller en kostnad. Övergången mellan ett tillstånd till ett annat är behäftat med osäkerhet, och beskrivs med en övergångssannolikhet. Markovegenskapen innebär att övergången bara beror av det aktuella tillståndet och vilken handling som avses, men inte av historiken.

Beslutsprocessen går ut på att maximera belöningen som erhålles när man kommer till ett nytt tillstånd. Belöningen kan vara positiv eller negativ och representerar hur önskvärt det är att befinna sig i tillståndet. Belöningen för respektive tillstånd samt övergångssannolikheterna antas vara kända. Att lösa en MDP innebär att hitta en optimal policy, som beskriver vilken handling som skall tas i varje steg för att uppnå maximal samlad belöning för hela kedjan av beslut. Policyer värderas och rankas utifrån deras förväntade värde efter ett bestämt antal beslutssteg, givet ett ursprungligt tillstånd. Detta görs genom att beräkna en värdefunktion som summerar det förväntade värdet, och maximera detta. Den policy som ger den maximala värdefunktionen kallas för en optimal policy.

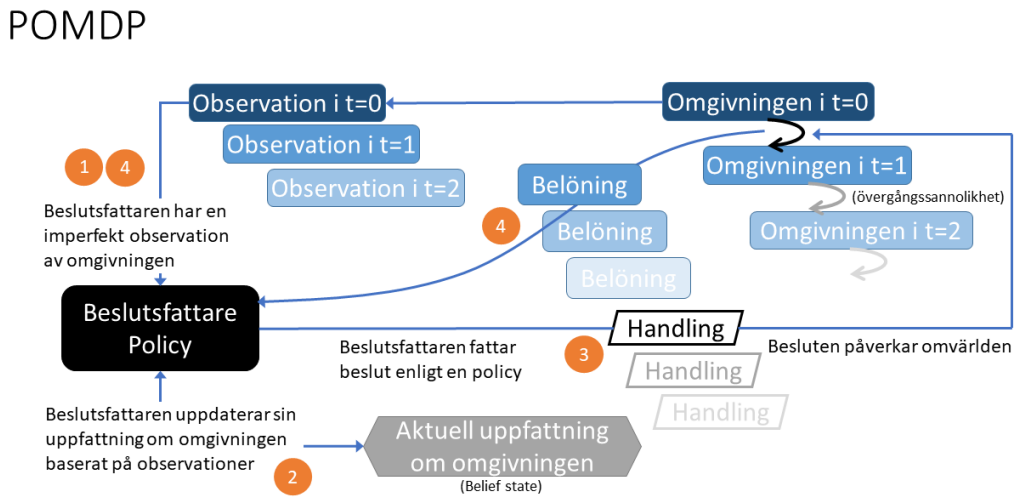
I MDP antas att ett systems tillstånd kan bli exakt identifierat, alltså att vi har all information vi behöver för att fatta ett beslut. När tillståndet är okänt eller delvis okänt är partiellt observerbara MDP:er ett sätt att angripa problemet.

2.2.2 Partially observable Markov decision process

I partiellt observerbara MDP:er (POMDP) är möjligheten att få en överblick över omgivningen (tillståndet) ofullständig eller omgivningen är endast delvis observerbart. Det går alltså inte att avgöra vilket tillstånd som råder, utan ett antal partiella observationer ger sannolikhetsinformation om vilket tillstånd som kan vara det aktuella. För att hantera detta

upprätthålls en lista på troliga tillstånd (trostillstånd, eng. belief state) som är en sannolikhetsfördelning över de möjliga tillstånden. Denna uppdateras i varje tidssteg baserat inte bara på senaste tillkommen information utan också utifrån tidigare observationer och handlingar. Sannolikhetsfördelningen över tillstånd ersätter därmed behovet av att lagra och beakta hela historien av observationer och handlingar, vilket i praktiken är ohanterligt. Det cykliska förfarandet illustreras i figur 3. Ett steg i cykeln är att en observation av omgivningen erhålls (1). Observationen används för att uppdatera (2) beslutsfattarens uppfattning omgivningen. Baserat på det väljer beslutsfattaren en handling att utföra (3). Den utförda handlingen resulterar i en belöning och en ny observation (4), och cykeln startar om.

POMDP är modeller som beskriver optimal styrning av dolda markovmodeller (hidden Markov models, HMM). Mängden handlingar för POMDP kan innehålla, förutom de för en MDP, även sådana som skapar fler eller säkrare observationer. Att lösa POMDP-problem är mer krävande än för MDP, och blir ofta ogörligt på grund av problemets höga dimensionalitet. Därför används i praktiken ofta approximativa lösningar. Värt att notera är att POMDP är en generellare klass av metoder än det mer kända området förstärkningsinlärning (eng. reinforcement learning).



Figur 3. I en POMDP vill man hitta en optimal policy som beskriver vilken handling som skall tas i varje steg för att uppnå maximal samlad belöning för hela kedjan av beslut. Handlingen påverkar omvärlden, men förändringen är behäftad med slumpmässighet som beskrivs med en övergångssannolikhet. Observationer av omvärlden används för att uppdatera uppfattningen (belief state), och denna uppfattning av omvärlden är det som guidar nästa beslut.

2.2.3 Forskningsläget

POMDP har funnit tillämpningar i en mängd olika fält. För att få en bild av var vi hittar tillämpningar för POMDP och vilka tekniker som används ges här en kort beskrivning utifrån de 50 högst citerade artiklarna i sökverktyget Web of Science de senaste 5 åren[7].

Ett stort område där POMDP:er används är inom navigering och ruttplanering för autonoma agenter så som förarlösa fordon, UAV:er, AUV:er och robotar, samt andra användningar inom robotik och människa-dator-interaktion. Ett annat tillämpningsområde handlar om att optimera olika typer av kommunikationsnätverk utifrån en mängd infallsvinklar, samt resursallokering för randteknik (eng. edge computing) och andra IoT-relaterade tillämpningar. Här finns även artiklar om optimering av elnät, samt olika problem avseende laddning av och kommunikation mellan självkörande fordon och UAV:er. Vidare finns det tillämpningar inom cyberattacker mot nätverk och smarta elnät, och motmedel mot utstörning av UAV-kommunikation. Planering och övervakning av underhåll av infrastruktur och andra tekniska system, samt energioptimering av byggnader är andra fält inom vilka

[7] Web of Science "Topic" = "POMDP" eller "partially observable Markov decision process"

POMDP används. Bland de högst citerade artiklarna återfinns några enstaka som handlar om tillämpningar inom sjukvård, modellering av biologiska system samt börshandel. Därtill kommer ett fåtal artiklar som utvecklar generella teorier och algoritmer för POMDP.

Den teknik som utforskas mest är djup förstärkningsinlärning, eller djup Q-inlärning. I Q-inlärning krävs ingen explicit modell av omgivningen, vilket är praktiskt när tillståndsrummet är stort som ofta är fallet. I detta problem lär sig modellen en approximation till den optimala funktionen Q för den optimala policyn, vilken svarar på vilken handling som skall tas i ett givet tillstånd. I praktiken används djupa neuronnät för att representera Q-funktionen.

Det har publicerats 170-tal artiklar om POMDP[8] per år de senaste åren. I det följande gör vi några nedslag utifrån den teoretiska utvecklingen bland de högst citerade artiklarna, samt några andra intressanta nedslag.

2.2.3.1 Kontinuerliga rum för tillstånd, handlingar och observationer

Det finns offline-metoder för att lösa små till medelstora POMDP:er, där hela eller stora delar av problemet löses innan exekvering av beslutskedjan. För att lösa stora problem behövs i allmänhet online-metoder, metoder där belief state och policyn uppdateras efter varje handling. Ett exempel på en online-algoritm är partiellt observerbar Monte-Carlo-planering (POMCP). Denna och andra algoritmer kan hantera kontinuerliga tillståndsrum. Andra algoritmer hanterar kontinuerliga handlingsrum, men det har varit mindre fokus på kontinuerliga observationsrum. En artikel av Sunberg och Kochenderfer (Sunberg och Kochenderfer 2018) tillhör de mest citerade artiklarna om POMDP de senaste åren, och presenterar två algoritmer som löser POMDP:er som har kontinuerliga rum för både tillstånd, handlingar och observationer. Den ena modellen kallas för “partially observable Monte Carlo planning with observation widening” och den andra “particle filter trees with double progressive widening”.

Monte Carlo-trädsökning (MCTS) är en effektiv algoritm för online-beslutsfattande. Trädet består av omväxlande tillståndsnoder och handlingsnoder, och genereras med slump men så att både bredd och djup utforskas. Detta sker bl.a. utifrån det uppskattade värdet för varje tillstånd-handlings-par. När tillstånds- och handlingsrummen är kontinuerliga kommer dock sannolikheten att utforska ett och samma tillstånd två gånger (vilket skapar djupet i trädet) vara noll, varför endast ett träd med djup 1 skapas. För att komma runt detta har en metod som kallas Double Progressive Widening (DPW) skapats. I denna införs en begränsning på antalet barn en nod kan ha, vilket tvingar fram en djupare utforskning.

POMCP är en metod för att komma runt den beräkningstunga bayesisk tro-uppdateringen, där historien från rotnoden till lövet lagras i varje nod. Detta har visat sig fungera bra för stora diskreta problem, men inte för kontinuerliga. I artikeln från Sunberg och Kochenderfer (Sunberg och Kochenderfer 2018) definieras en algoritm som kombinerar POMCP med DPW, för att sedan teoretiskt bevisa att den leder till en suboptimal lösning (en lösning som antar att problemet blir fullt observerbart efter ett tidssteg). Den första av två algoritmer som presenteras som lösning på detta går ut på att införa vikter på tro-uppdateringarna i POMCP. Varje observationsnod innehåller en samling besökta *viktade* tro-tillstånd, och ett av dessa samplas för att välja nästa handling.

Den andra algoritmen baseras på MCTS och DPW. Istället för att skapa en trädsökning i tillståndsrummet, tar man fram lösningen med avseende på sekvenser i tro-rummet. Partikel-filtrering används för att ta fram den beräkningssvåra bayesisk tro-uppdateringen. Simulerade experiment på ofta använda jämförelseproblem används för att utvärdera algoritmerna, vilket visar på en blandad framgång. Flera av de beskrivna problemen är dock för enkla för att lyfta fram styrkan i de föreslagna algoritmerna.

[8] Enligt sökverktyget Web of Science (www.webofscience.com) (besökt juni 2023)

En annan högt citerad artikel som behandlar samma problem, nämligen att ha kontinuerliga tillståndsrum, handlingar och observationer, är publicerad av Jiang m.fl. (Jiang m.fl. 2019). De noterar att många lösningar bygger på diskretisering av problemet, men det är problematiskt då dimensionaliteten typiskt blir mycket stor och prestandan låg. Därför utvecklas istället en simulerings-baserad algoritm för att hitta den lokalt optimala policyn för en POMDP givet observationer, baserad på en kontinuerlig formulering av problemet. Algoritmen är inte begränsad av några antaganden, och behöver inte heller någon information på förhand. Därför kan den appliceras på de flesta beslutsproblem med kontinuerliga rum.

Problemet löses i två steg. Först utvecklas matematiska förutsättningar innefattande gradienten av *prestandapotentialen* med avseende på policyn. Prestandapotentialen mäter den förväntade belöningen givet en viss observation och en policy. Två satser för att beräkna gradienten presenteras, en för slumpmässig policy och en för deterministisk policy. Dessa bygger på känslighetsanalys, och visar på att gradienten med avseende på slumpmässiga policyer kan uppskattas utifrån en enstaka följd av handlingar och observationer, utan på förhand given information om funktionerna för tillståndsövergångar, observationer eller belöning. Ett antagande är att markovkedjan är ergodisk, dvs. att dess egenskaper kan skattas från en enskild sekvens av beslut. Detta gäller i fallet slumpmässig policy. Författarna hävdar att detta ofta gäller.

Den optimala policyn kan erhållas utifrån gradienterna. Dessa är emellertid praktiskt svårlösliga. I det andra steget utvecklas därför en simuleringsbaserad iterativ algoritm, med låg beräkningskomplexitet. Denna bygger på approximationer till de framtagna ekvationerna, och itererar över policyer. Endast gradienten av policyfunktionen och den ackumulerade belöningen behövs i varje iteration. Det visas att den föreslagna algoritmen konvergerar. Numeriska beräkningar visar på förbättringar i prestanda jämfört med två existerande diskretiseringsbaserade lösningar.

2.2.3.2 Förklarbarhet

Chadès och medförfattare (Chadès m.fl. 2021) noterar att visuell representation och förståelse av den optimala lösningen (policyn) är viktiga komponenter för att beslutsfattare skall ha tillit till beslutspolicyer som är framtagna av MDP/POMDP-modeller. Ökad förklarbarhet av MDP-problem kan ses som XAI (Ferrer-Mestres m.fl. 2021).

I många problem kan tillståndsrummet bli så stort att det blir oöverblickbart (Ferrer-Mestres m.fl. 2020). Metoder för att abstrahera tillstånd syftar till att reducera storleken på stora tillståndsrum genom att aggregera tillstånd som är liknande utifrån något mått. Tillståndsabstraktion har i huvudsak fokuserat på att hitta det minsta tillståndsrummet för en given reduktion i prestanda.

Ferrer-Mestres m.fl. (Ferrer-Mestres m.fl. 2020) presenteras en ny lösning (”K-MDP”) där reduktionen i prestanda minimeras utifrån ett givet maximalt antal tillstånd (K). På så sätt kan K väljas så att problemet blir greppbart för en mänsklig beslutsfattare, men ändå lösas optimalt. Utifrån en ursprunglig MDP (med många tillstånd) samt begränsningen (K) på antalet önskade tillstånd, genereras en MDP med ett mindre tillståndsrum sådan att skillnaden mellan den ursprungliga värdefunktionen och den nya värdefunktionen minimeras.

Det finns i huvudsak två sätt att visualisera en POMDP-lösning; genom en graf som representerar policyn eller ett diagram med vektorer som visar värdefunktionen i olika tillstånd. Värdefunktionen är styckvis linjär, och kan modelleras med en ändlig mängd linjära funktioner, så kallade α-vektorer. Varje α-vektor är associerad till en viss handling (Chadès m.fl. 2021).

Som nämnts är policy-grafer för realistiska problem alldeles för komplexa för att vara överblickbara för en mänsklig beslutsfattare (Chadès m.fl. 2021; Ferrer-Mestres m.fl. 2021). För att komma tillrätta med detta har Ferrer-Mestres och medförfattare utvecklat en metod som kraftigt minskar antalet tillstånd (K) och antalet α-vektorer (N), som de kallar K-N-MOMDP (Ferrer-Mestres m.fl. 2021). De visar på beräkningar som genererar mer

kompakta och förståeliga policy-grafer, samtidigt som de bibehåller liknande värde. Det ena problemet löses med ovan nämnda metoder att abstrahera tillstånd, medan tre nya algoritmer presenteras för att lösa problemet med α-vektorerna. MOMDP står för Mixed Observable MDPs och är sådana där delar av tillståndsrummet är fullständigt observerbart medan andra delar endast är delvis observerbara (POMDP) (Ong m.fl. 2010).

2.2.3.3 Kunskapsbaserad belöning

I problem där aktiv inhämtning av information för att erhålla en bättre bild av omgivningen (vissa tillståndsvariabler) är den ursprungliga formuleringen av POMDP inte lämplig eftersom den är formulerad så att belöningen bara beror på tillståndet och handlingen, och alltså inte fångar någon information om kunskap från tidigare utforskande. I sådana situationer så som där övervakning och utforskande är viktigt är det mer relevant att beakta belöning baserad på erhållen kunskap, representerad av de troliga tillstånden (belief states) istället. För att hantera detta introducerar Araya-López (Araya m.fl. 2010) ρ-POMDP, där belöningen omformuleras till att vara väntevärdet av belöningen utifrån belief state. I artikeln från 2010 visas på matematiska lösningar för olika fall, specifikt när värdefunktionen är styckvis linjärt komplex, samt approximationer när den inte är det.

I flera fall är tro-beroende belöningen ρ inte konvex, och för dessa kan inte problemet lösas med tidigare föreslagna metoder (Fehr m.fl. 2018). Fehr och medförfattare (Fehr m.fl. 2018) visar att för ρ-POMDP:er med λ_ρ-Lipschitz belöningsfunktion och ändliga horisonter så är optimala värdefunktionen fortfarande Lipschitz-kontinuerlig (begränsad i hur snabbt den kan ändras, så att för alla val av två punkter på kurvan är absolutvärdet av lutningen mellan punkterna begränsad till en konstant) istället för styckvis linjärt konvex. En utökad definition av Lipschitz-kontinuitet används, där Lipschitz-konstanten ersatts av en vektor, samt lokal kontinuitet snarare än uniform antas. Utifrån detta utvecklas approximationer till den optimala värdefunktionen som är bundna både uppåt och nedåt.

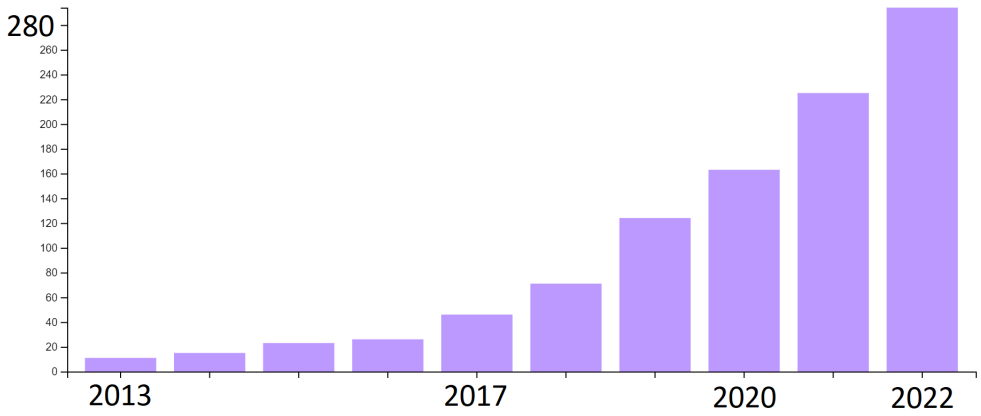
2.3 Framtidsprognos och analys

Multikriterieanalys (MCDM) är en effektiv teknik för att välja ett beslut utifrån flera alternativ. När detta grundar sig i data och beslutet leder till handling utgör MCDM en metod för preskriptiv dataanalys. Det finns många metoder inom området och nya utvecklas och tillämpas på en mängd olika problemområden. Dynamiska omständigheter kan vara något som är viktigt i ett militärt sammanhang, där beslut skall fattas i en omgivning i snabb förändring, varför det kan vara relevant att titta på dynamiska beslutsmetoder. Mer statiska metoder kan dock fungera på strategisk nivå. POMDP är en mogen metod som används inom en mängd områden, och det finns en mängd sätt att lösa POMDP på. Trots det finns det fortfarande problem kvar att lösa, och det görs fortfarande teoretiska landvinningar. Vikten av de verk som presenterats ovan beror på vilken tillämpning som avses, POMDP kan användas för en mängd problem, bl.a. för beslutsfattande hos autonoma agenter men också för att värdera handlingsalternativ vid mänskligt beslutsfattande. Tankarna om förklarbarhet kan vara av speciell vikt vid användande av metoden för militärt beslutsfattande där människan är med i beslutsloopen.

3 Osäkerhetshantering

Dataanalys (DA) är ett centralt begrepp inom en rad olika närbesläktade områden såsom artificiell intelligens (AI), maskininlärning och data science, som alla syftar till att designa och träna modeller med hjälp av data som sedan kan användas i olika former av beslutssammanhang kring någon verklighet av intresse. Framförallt bygger den senaste tidens popularitet och användning av AI i mycket hög utsträckning på framsteg inom DA, särskilt inom ett område kallad djupinlärning (LeCun, Bengio, och Hinton 2015). Denna ansats till DA är baserad på en viss typ av modell som fångar upp mönster på olika abstraktionsnivåer i en hierarki. I sin ursprungsform har dock inte dessa typer av modeller någon explicit representation av osäkerhet, t.ex. i form av en sannolikhetsfördelning, vilket innebär att utdata från modellerna, såsom prediktering av en klass, inte är baserat på en sådan representation även om man kan normalisera resultatet så att det formellt uppfyller kravet rent matematiskt på en sannolikhetsfördelning. Med en alltmer frekvent användning av AI och djupinlärning inom en rad olika tillämpningar i samhället har det dock uppmärksammats ett behov av sådan modellering (Abdar m.fl. 2021). Detta för att tillämpningar ofta är av en art som innehåller osäkerhet som en naturlig komponent såsom beslut rörande någon dynamisk miljö med osäkerhet associerad med nuvarande situation samt, således, också framtida situation inklusive möjliga konsekvenser av beslut. Dessutom har i samband med den alltmer utbredda användningen av djupinlärning och AI inom olika tillämpningar inom samhället uppstått ett ökat tryck på grundläggande förståelse hur dessa modeller fungerar, benämnt förklarbar AI (eng. explainable AI/XAI[9]) (Arrieta m.fl. 2020). Grundtesen är att om AI ska kunna användas i viktiga samhällsfunktioner där någon aktör bär ansvar så måste också AI vara förklarbar i termer av hur ett beslut eller klassificering har gjorts. Detta medför också ett indirekt krav på att modeller ska kunna hantera och representera osäkerhet inom DA om de är tänkta att användas i en miljö där det existerar explicit osäkerhet och kanske framförallt om det också finns kopplad någon typ av riskfaktor till sådan osäkerhet. En sån riskfaktor kan röra att ett beslut fattas baserat på en osäkerhet i situationen som potentiellt skulle kunna leda till allvarliga negativa konsekvenser, såsom i säkerhetskritiska tillämpningar.

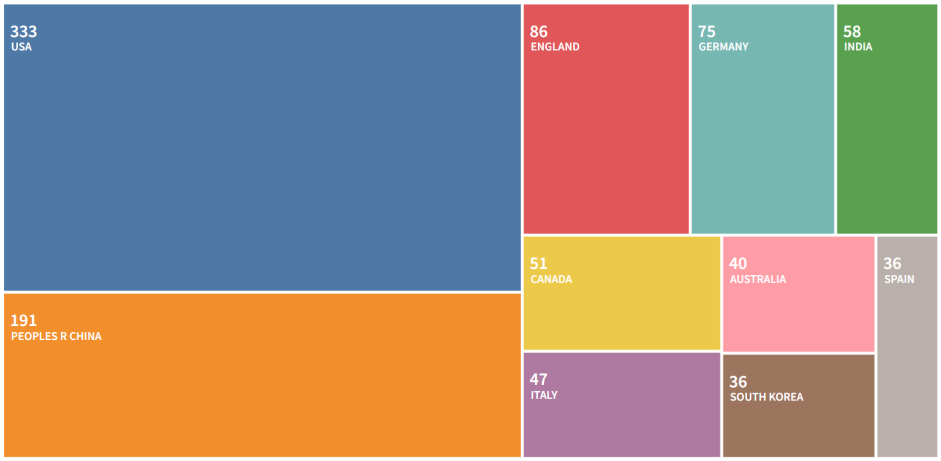
En sökning av publicerade artiklar med Web of Science innehållandes orden ”uncertainty” och ”machine learning” i titel, sammanfattning eller nyckelord gav drygt 1000 träffar under perioden 2012-2022.[10] Som framgår av figur 4 har intresset ökat stadigt under de senaste åren. I figur 5 visas forskningens ursprungsländer.



Figur 4. Staplarna visar antal publicerade artiklar, år för år, mellan 2012 och 2022, som nämner ”uncertainty” och ”machine learning” i titel, sammanfattning, eller nyckelord

[9] Se även kapitel 4

[10] WoS-sökfråga: (TS=(uncertainty NEAR machine learning) AND (SU = (Computer Science) OR SU = (Mathematics) OR SU = (Automation & Control Systems)OR SU = (Engineering)OR SU = (Remote Sensing) OR SU = (Robotics) OR SU = (Science & Technology Other Topics) OR SU = (Telecommunications))) AND PY=(2012-2022)



Figur 5. Forskningens ursprungsländer för söktermen ”uncertainty NEAR machine learning” 2013–2022

3.1 Osäkerheter inom dataanalys

Modellering av osäkerhet har en lång tradition av forskning inom AI-området. En ofta förekommande uppdelning av olika typer av osäkerheter är inom de två kategorierna aleatorisk och epistemologisk osäkerhet som också många gånger benämns som oreducerbar respektive reducerbar osäkerhet (Oberkampf m.fl. 2004; Ferson m.fl. 2004). Vilka osäkerhetsrepresentationer eller ramverk som har förmåga att verkligen modellera dessa olika typer råder det en viss oenighet kring. Som ett exempel finns det argument för att sannolikhetsfördelning inte är en tillräcklig rik matematisk struktur för att i det generella fallet kunna uttrycka båda typer av osäkerhet och därför har alternativa representationer utvecklats (Walley 2000), t.ex. genom att uttrycka sannolikheter som intervall istället för ett exakt kontinuerligt värde.

Inom dataanalys, har de två typerna av osäkerhet, dvs. *aleatorisk* och *epistemologisk*, översatts till olika delar som förekommer inom sådan analys (Hüllermeier och Waegeman 2021). En klar uppdelning mellan dess delar är dock svår att göra då osäkerhet inom en del ofta överförs till en annan. Som ett exempel så överförs osäkerhet genom att man använder en approximativ träningsalgoritm, vilket snarare är mer regel än undantag på grund av komplexiteten i moderna datamängder/modeller, explicit eller implicit till prediktionssteget då man använder modellen. Avgörande för förmågan att representera olika typer av osäkerheter faller på valet av modellstruktur, träningsansats samt förhållande dessa emellan.

3.1.1 Modellstruktur

Modellstruktur och dess komplexitet, (t.ex. dimension och antal nivåer i en hierarki) sätter en ram för hur väl olika komplexa samband kan modelleras, vilka osäkerheter som kan fångas upp samt vilka träningsansatser som är mer eller mindre lämpliga. Modellstrukturen är också avgörande för efterföljande analys, t.ex. kopplat till förklaringsbarhet (Arrieta m.fl. 2020). En tydlig skiljelinje finns här mellan modeller tillhörande området djupinlärning (LeCun, Bengio, och Hinton 2015) kontra modeller baserat på statistisk modellering (Gelman m.fl. 2013) då den första syftar till att fånga upp mönster genom en generell modellstruktur som inte är direkt kopplad till någon slags grundläggande design av struktur kring hur data har genererats medan man vid den andra ansatsen gör antaganden, samt designar, hur data genererats genom sannolikhetsfördelningar[11].

[11] I fortsättningen av kapitlet skriver vi i vissa fall förkortningen fördelning istället för sannolikhetsfördelning.

3.1.2 Träningsansats

Modellstrukturen samt mängden data påverkar i hög grad träningsansatsen som i de flesta fall innebär att identifiera parametrar i modellen som på något sätt matchar hur data genererats. Inom den bayesisk statistiska modelleringen (Gelman m.fl. 2013), som är den mest förekommande ansatsen inom AI och maskininlärningsområdet för osäkerhetsmodellering, fångar träningsansatsen osäkerhet kring de ovan nämnda parametrarna genom fördelningar men till kostnad av ökad komplexitet. Detta resulterar i en övervägande majoritet av tillämpningar behöver använda approximativa algoritmer. Detta leder även till en ytterligare nivå av osäkerhet, dvs. hur bra approximation genererar en sådan algoritm med avseende på den verkliga fördelningen över parametrar? Valet av vilken typ av approximationsalgoritm man använder beror på mängden data man vill använda för träning men även typ av modellstruktur då det finns svårigheter för träningsansatser att hantera vissa strukturer. Det finns flera olika typer av träningsansatser (inkluderat även ansatser som inte direkt syftar till att fånga upp osäkerhet): de som bygger på optimering, ofta förekommande inom djupinlärning, dvs. identifiera parametrar utifrån ett optimeringsproblem kopplat till data och målfunktion (LeCun, Bengio, och Hinton 2015) samt approximativa algoritmer som bygger på att man kan sampla från målfördelningen (M. D. Hoffman och Gelman 2014). De finns även mer komplexa ansatser som bygger på att man approximerar en målfördelning med en enklare fördelning som man sedan använder för att identifiera parametrar, så kallad *variational inference* (VI) (Blei, Kucukelbir, och McAuliffe 2017). Generellt sett är ansatser som bygger på optimering mindre resurskrävande och används därför mer frekvent vid komplexa modeller och stora datamängder, t.ex. djupinlärning, medan samplingsbaserade metoder används i fall där det är rimligt utifrån problem och data. Vissa metoder inom optimering, t.ex. VI, fångar upp osäkerhet men med en viss oklarhet kring vilken garanti man får i termer av likhet med den verkliga fördelningen (Yao m.fl. 2018) medan de samplingsbaserade metoderna (ofta benämnda ”Markov-Chain Monte-Carlo” eller liknande) har mer av teoretiska garantier givet att ett antal kriterier är uppfyllda vilket i sig kan vara svårt att påvisa (Betancourt 2017).

3.2 Forskningsläget

Vi beskriver forskningsläget utifrån två dominerande, men olika, områden som återfinns inom maskininlärning / AI / DA, nämligen: djupinlärning, som återfinns i avsnitt 3.2.2 och statistisk maskininlärning som tidigare nämnts är präglad av den bayesiska ansatsen till statistik (Gelman m.fl. 2013) och som återfinns i avsnitt 3.2.3. Delområdena inom respektive område är utvalda utifrån vad som varit och fortsatt är områden där mycket forskning pågått / pågår och som på olika sätt tar sig an problemet med att representera osäkerhet.

En intressant iakttagelse är att de två områdena som har ganska olika ansatser till dataanalys alltmer börjar bli sammanflätade, framförallt så används bidrag från statistisk maskininlärning till stor grad i metoder för djupinlärning som ett sätt att uppnå osäkerhetsrepresentation. Detta syns t.ex. inom delområdet bayesisk djupinlärning, beskrivet i avsnitt 3.2.2.1. En annan tidigare framgångsrik ansats inom maskininlärning som nu även anammats inom djupinlärning, beskrivit i avsnitt 3.2.2.2, är ensembletekniken som går ut på att man använder en mängd olika modeller och på sätt skapar bättre prediktionskvalité. Till sist inom djupinlärningsområdet så beskrivs en ansats som går ut på att erhålla en viss typ av tolkning av sannolikheter som resultat från djupinlärningsmodeller, beskrivet i avsnitt 3.2.2.3, samt några andra ansatser inom avsnitt 3.2.2.4.

I avsnitt 3.2.3.1 tas ett återkommande problem inom statistisk maskininlärning upp, nämligen träningsalgoritmens beräkningskomplexitet och skalbarhet. Detta följs upp med avsnitt 3.2.3.2 som handlar om en ökad grad av automation som möjliggör att man utforskar modeller utan att stora härledningsarbeten vad gäller matematiska uttryck osv. behövs utföras manuell. Till sist, i avsnitt 3.2.3.3, beskrivs en rik modell ur representationssynpunkt som också blivit alltmer sammanflätat med djupinlärning, nämligen gaussiska processer. Detta följs till sist, i avsnitt 3.2.3.4, av några andra viktiga ansatser inom området.

3.2.1 Källor

Utvecklingen inom området har en snabb takt och drivs i stort av ett mindre antal välrenommerade konferenser samt i viss utsträckning tidskrifter. Efterföljande text beskriver främst forskningsläget utifrån dessa forum:

- Advances in Neural Information Processing Systems (NeurIPS), konferens
- Journal of Machine Learning Research (JMLR), tidskrift
- Proceedings of Machine Learning Research, samlar ett stort antal proceedings från välrenomerade konferenser/workshops inom området maskininlärning

Utöver dessa har ett antal andra artiklar, t.ex. i form av litteraturstudier kring ett visst delområde, använts på olika sätt t.ex. för att få en översikt av delområden såsom Bayesian deep learning, men också som ett sätt att vidare utforska vissa nyckelreferenser. Sökning, främst inom tidsperioden 2012 – 2022 med en viss viktning mot den senare delen då utvecklingen inom området har en hög takt, har skett genom en kombination av Google Scholar för att erhålla citeringsinformation samt genom sökning på forumens egna webbsidor och vidare har även citerat material i artiklar följts upp. För att fånga upp bidrag som är relaterat till någon form av dataanalys med relation till osäkerhet har sökfrågor utformats som antingen fångar upp en generell metodik för modellering, såsom ”Bayes”[12], eller generell metodik som behövs för att träna en modell, såsom ”Markov chain” och ”variational”. Detta resulterade i sökningar med termer såsom ”Bayes”, ”uncertainty”, ”variational” och “Markov chain" men även mer detaljerade termer kan ha använts vid vidare utforskning av en viss specifik metod. Utveckling över tid i forumen har studerats manuellt utifrån tidigare erfarenhet inom området. Vidare har även publikationer vid arXiv använts där typen av bidrag är blandad, t.ex. en del kan vara preprints eller liknade till publicerat material vid annan tidskrift medan andra bidrag kan ligga där i väntan på att granskningsprocesser färdigställts.

3.2.2 Djupinlärning

Djupinlärningsmetoder är baserade på olika typer av hierarkiska arkitekturer där främst neuronnät används som grundkomponent (LeCun, Bengio, och Hinton 2015). Idén bakom dessa olika arkitekturer är att fånga upp mönster på olika abstraktionsnivåer och på så sätt erhålla bättre prediktiv kvalité. Proceduren för att erhålla en färdigställd djupinlärningsmodell bygger på att man först definierar en struktur i termer av antal lager i hierarkin samt antal noder, så kallade neuroner, på varje nivå som sedan också bestämmer komplexiteten av modellen. Ju fler lager och noder desto mer komplex modell som kan fånga upp ett större antal mönster men där man också i större utsträckning riskerar att överanpassa (eng. overfitting) modellen till data som då får som följd att modellen generaliserar dåligt vilket innebär låg prediktiv kvalité. Efter arkitekturval så anpassar man, ofta benämnt tränar, modellen till data vilket kan t.ex. innebära att man försöker identifiera de parametrar, vikter, som återfinns i modellen och som passar bäst gentemot någon målfunktion som är baserat på det sanna/korrekta värdet av en målvariabel.

3.2.2.1 Bayesisk ansats till djupinlärning

Traditionellt har djupinlärningsmodeller tränats genom optimering för att identifiera en uppsättning av parametrar som sedan då används i en viss tillämpning för att prediktera något av intresse. Ett problem med denna ansats är att man inte direkt modellerar osäkerheten vad gäller dessa parametrar och att man därför inte fångar upp den osäkerhet som egentligen existerar i en viss given prediktion. För att belysa denna problematik har det sedan 2016 hållits en workshop benämnd, bayesisk djupinlärning (eng. Bayesian Deep Learning) vid ett utav det främsta forumet för djupinlärning, ”Advances in Neural Informat-

[12] Alla dessa termer kommer att sättas i en kontext samt förklaras i efterföljande text.

ion Processing Systems” (NeurIPS), som speciellt fokuserat på hur man kan träna djupinlärningsmodeller utifrån ett bayesiskt perspektiv. Terminologin för bayesisk djupinlärning är inte helt stabil då det också kan benämnas bayesiska neuronnät (Gawlikowski m.fl. 2021) men termen kan även användas i ett sammanhang av mer specifik betydelse (Wang och Yeung 2020).

Forskningsläget inom bayesisk djupinlärning visar på en bredd inom framförallt olika ansatser för att träna en modell på ett sätt som i viss mån återger posteriori distributionen som är målet för den bayesiska ansatsen (Abdar m.fl. 2021). Ansatserna bygger på olika varianter, ofta som en konsekvens av att hantera komplexitet, av de huvudsakliga ansatserna som nämndes under avsnitt 3.1.2, såsom varianter av, stochastic gradient descent (SGD) (Maddox m.fl. 2019), variational inference (VI) (M. D. Hoffman m.fl. 2013; Gal och Ghahramani 2016), samt Hamiltonian Monte-Carlo (HMC) (M. D. Hoffman och Gelman 2014; T. Chen, Fox, och Guestrin 2014). Ansatserna syftar till att på olika sätt approximativt fånga upp posteriori fördelningen, eller någon typ av osäkerhet, över de vikter som återfinns i neuronnät. Då själva djupinlärningsmodellen har egenskaper, såsom t.ex. multimodalitet, så finns det tveksamheter till i vilken utsträckning många av dessa metoder verkligen fångar upp all den osäkerhet som finns över parameter rymden. Som ett exempel kan nämnas att VI ofta använder fördelningar av enkelmodal karaktär (ett optimum eller ”topp” i fördelningen) som då kan ge en förenklad vy av den verkliga (posteriori) fördelningen över parameterrymden (Izmailov m.fl. 2021). Problemet liknar det klassiska optimeringsproblemet inom där algoritmer kan fastna i lokala optimum. Värt att nämna i sammanhanget är att den träningsansats, HMC, som anses vara ”state-of-the-art” inom bayesisk dataanalys och som är beräkningstung kan vara besvärlig att använda för att navigera i en parameterrymd som utgörs av neuronnät. Detta är i högsta grad ett aktivt forskningsområde och en nyligen gjord studie (Izmailov m.fl. 2021) visar intressanta till mesta dels fördelaktiga resultat från träning med HMC jämfört med andra typer av approximationer. Studien påvisar dock också problematik kring konvergens inom parameterrymden men dock inte prediktionsrymden (benämnt funktionsrymd i studien) vilket i stort är det man är intresserad av inom maskininlärningsområdet (där en metod för prediktion är i fokus snarare än att estimera okända parametrar i någon modell). Ett annat spännande resultat från den studien är även att de prediktionsfunktioner man erhåller som så kallade exempel (eng. samples) kan vara relativt olika vilket är något som då i praktiken står i kontrast till de metoder som är mer av karaktären optimering kring ett enskilt optimum.

En av de senare utvecklingarna inom området som har belysts (Tran m.fl. 2022; Fortuin 2022) är valet och vikten av att mer försiktigt välja, eller snarare designa den, så kallade priori fördelningen som utrycker en fördelning över parametrar innan man använder träningsalgoritmen för att anpassa modellen till data. I en översiktsartikel (Fortuin 2022) konstateras att ofta väljs apriori-fördelning utifrån ett perspektiv av icke-informativ fördelning genom normalfördelningen. I praktiken, då djupinlärningsmodeller är mycket högdimensionella och innehar komponenter av icke-linjär natur, så kan det vara mycket svårt att överskåda vilken typ av funktionsrymd kombinationen av sådana icke-informativa fördelningar genererar (Tran m.fl. 2022).

3.2.2.2 Djupa ensembler

En ansats som belysts av litteraturstudier kring osäkerhetshantering inom djupinlärning (Ganaie m.fl. 2022; Gawlikowski m.fl. 2021; Abdar m.fl. 2021) men även genom ett flertal andra litteraturstudier (Dong m.fl. 2020; Sagi och Rokach 2018) av bredare karaktär är en metodik/teknik inom maskininlärning som benämns ensembler. Huvudidén med ensembler är att man på olika sätt slår ihop resultat från ett antal olika modeller och på så sätt får ett bättre resultat i termer av prediktion. Konceptet bygger på att medlemmarna i ensemblen överlag är bra på att prediktera men att de är olika, framförallt att de gör olika typer av fel, vilket då medför en förbättring vid sammanslagning av resultat. Det kan sägas att tekniken i sig inte har ett explicit sätt att modellera osäkerhet men implicit så uppnås detta genom att varje medlem i ensemblen kan göra olika typer av prediktioner vilket då utgör en typ av implicit osäkerhetsrepresentation.

Ensembler där medlemmarna utgörs av djupinlärningsmodeller har belysts och utvärderats från flera perspektiv (Wilson och Izmailov 2020; Lakshminarayanan, Pritzel, och Blundell 2017; Gustafsson, Danelljan, och Schon 2020) där intressanta observationer gjorts kring sambandet mellan den bayesiska metodiken för djupinlärning och dess osäkerhetsrepresentation för prediktion samt den implicita osäkerhetsrepresentationen som uppstår genom ensemblemodellering. Experiment (Wilson och Izmailov 2020) visar på en hög grad av likhet mellan osäkerhetsrepresentation av dessa typer av metodiker vilket kan förefalla märkligt då ensembler inte har någon egentlig tydlig koppling till den bayesiska ansatsen till dataanalys men som har belysts (Gustafsson, Danelljan, och Schon 2020) så är en av själva grundidén med både ensembler och bayesisk metodik att en modellrymd utforskas och används varvid likhet uppstår. Intressant nog finns det resultat som visar att ensembler till och med kan generera resultat närmre mer resurskrävande bayesisk analys än ett flertal andra mer uttalade approximativa bayesiska metoder (Wilson och Izmailov 2020; Gustafsson, Danelljan, och Schon 2020; Izmailov m.fl. 2021). En problematik med ensemblemetoder kan vara att de är beräkningsmässigt mer kostsamma då tekniken bygger på att man skalar upp antal tränade modeller med en faktor som är antalet medlemmar i ensemblen.

3.2.2.3 Kalibreringsmetoder

Kalibreringsmetoder (se översikt (Gawlikowski m.fl. 2021)) bygger på en idé från ett frekventistiskt perspektiv (Lakshminarayanan, Pritzel, och Blundell 2017) att modeller som ger ifrån sig ett resultat i termer av sannolikheter ska relatera till modellens förmåga att prediktera korrekt klass utifrån att sannolikheter uttrycker förekomster av något i termer av relativ frekvens (andel / utfallsrum). Om en modell ger en sannolikhet p för en viss klass för ett givet antal instanser som ska predikteras så bygger kalibreringsidén på att $p \cdot 10^2$ % av dessa instanser har den givna klassen (Minderer m.fl. 2021) vilket då kallas att modellen är kalibrerad eftersom sannolikheten då matchar klassisk sannolikhetsteori i termer av relativ frekvens. Fördelen med detta är att man direkt kan se utifrån ett givet resultat från en modell hur tillförlitlig en klassificering är, vilket då kan användas på olika sätt i olika tillämpningar. Det finns olika metodiker för kalibrering, t.ex. att utjämna sannolikheter så att ovanstående egenskap erhålls (Guo m.fl. 2017), som antingen kan utföras som en del av träning eller som en metodik efter träning (se vidare (Gawlikowski m.fl. 2021)). En problematik som belysts är att ett av de mer vanligt förekommande måtten för kalibrering behöver beräknas genom en procedur som resulterar i en icke kontinuerlig struktur i form av en uppdelning i diskreta intervall och beroende på antal sådana intervall så kan olika resultat erhållas (Minderer m.fl. 2021). Att utreda när och varför en viss modell är kalibrerad är ett något snårigt område men det finns empiri som talar för att man kan relatera hur väl en modell generaliserar till graden av kalibrering (Carrell m.fl. 2022).

3.2.2.4 Övriga ansatser

Förutom att modellera osäkerhet kring parametrar såsom i den bayesiska ansatsen till djupinlärning så finns det även andra ansatser till osäkerhetsmodellering inom området som bygger på att man istället tränar en djupinlärningsmetod att ge som resultat en viss given osäkerhetsstruktur. Traditionellt inom statistik och andra sammanhang där osäkerhet ska modelleras så har sannolikhetsfördelningar använts men det finns även viss kritik att dessa strukturer saknar förmåga att särskilja mellan de två huvudtyperna (se tidigare avsnitt 3.1) aleatorisk och epistemologisk osäkerhet (Walley 2000). Som ett sätt att hantera detta finns det ansatser som syftar till att lära sig en mer komplex struktur (Sensoy, Kaplan, och Kandemir 2018; Manchingal och Cuzzolin 2022).

En annan viktig aspekt inom området som särskiljer sig från statistisk maskinlärning är att själva modellarkitekturen, vilket då också är en typ av osäkerhet, för en djupinlärningsmodell är svår att designa utifrån givna välmotiverade regler. Detta leder till att man inom djupinlärningsområdet ofta behöver utföra olika typer av experiment innehållande olika arkitekturer med ibland ett antal något godtyckliga val tills man erhåller ett resultat som

har önskvärda utvärderade egenskaper. Ett sätt att ta sig an arkitekturproblemet är att definiera problemet som ett sökproblem, benämnt neural arkitektur sökning (eng. neural architecture search) och på olika sätt automatisk söka av olika möjligheter för arkitekturer, se vidare (Elsken, Metzen, och Hutter 2019; Ren m.fl. 2021).

3.2.3 Statistisk maskininlärning

Statistisk maskininlärning är i, konstrast till djupinlärningsmodeller, baserade på någon form av design, utifrån data eller problem, av modell baserat på sannolikhetsfördelningar. Utgångspunkten för designen kan dock vara olika, antingen att man innehar någon typ av domänkunskap kring processen där data har genererats och att man bygger upp modellen utifrån det, t.ex. att man sedan innan vet att data följer en normalfördelning (som ett typexempel). Design kan också utgå mer utifrån ett explorativt perspektiv eller utifrån ett beslutsperspektiv såsom att definiera en typ av mönster via fördelningar som ska fångas upp och som man förmodar kan inneha en slags nyttoeffekt för beslutsproblemet i termer av att det korrelerar med någon annan aspekt av problemet eller till mänsklig förståelse, t.ex. att fånga upp kluster av datapunkter via en mix av fördelningar. Området domineras av den bayesiska ansatsen till statistisk inferens, så kallad bayesisk inferens (Gelman m.fl. 2013), där huvudansatsen till tillämpningsnära problem blir att använda olika typer av approximativa algoritmer för att få en estimering av posteriorifördelningen över de ingående parametrarna i modellen och som då kan användas för prediktion. Anledningen till det stora fokuset på approximativa algoritmer är att slutna analytiska uttryck många gånger inte kan härledas eller att det innebär för stora begränsningar till de olika designval som ingår i modellen (t.ex. val av prior).

3.2.3.1 Skalbarhet

En kärnproblematik inom bayesisk dataanalys som har varit och fortfarande är ett utav de främsta problemen som forskare tar sig an är beräkningsbarhet av posteriori fördelning vilket är målet för dataanalysen. Som beskrevs i avsnitt 3.2.2 innebär osäkerhetsmodellering inom djupinlärning att de problem som återfinns inom bayesisk dataanalys i stort överförts till djupinlärning. Områdena har alltså blivit alltmer sammanflätade. Problemet kvarstår i hög grad på grund av att datagenereringstakten i samhället har ökat, vilket gör att problem som kunde lösas tillfredsställande med en viss metod i dåtid relativt snabbt blir utdaterat med dagens datamängder. Det kan också sägas att problemet att beräkna en posteriorifördelning är mer komplex än att t.ex. använda optimeringsbaserade metoder, t.ex. SGD (LeCun, Bengio, och Hinton 2015), som ofta är utgångspunkt inom djupinlärning. Det finns dock optimeringsansatser som är specialutvecklade för stora datamängder som bygger på VI (se vidare avsnitt 3.1.2) (M. D. Hoffman m.fl. 2013) men det kan vara problematiskt att utvärdera kvalitén av sådana resultat (Yao m.fl. 2018). På grund av detta så är det ofta intressant att använda samplingsbaserade metoder då de kommer med vissa teoretiska garantier om konvergens uppnås (vilket i sig kan vara svårt att visa). Dessa metoder är dock i sin traditionella form beroende av att gå igenom datapunkter ett flertal gånger per erhållen samplad datapunkt från posteriori fördelning (N. Chen, Xu, och Campbell 2022). För att komma tillrätta med det här problemet så finns det i huvudsak två ansatser (Bardenet, Doucet, och Holmes 2017): (1) dela upp och parallellisera beräkningar (2) reducera antal datapunkter med bibehållna egenskaper utifrån hela datamängden. Båda (1) och (2) innehåller olika typer av utmaningar kring att de approximativa algoritmerna inte tar del av hela datamängden vilket kan innebära problem i högdimensionella rymder (Betancourt 2015). För (1) så är ett kritiskt steg hur man aggregerar informationen från de olika delarna och olika ansatser för detta återfinns i litteraturen (Minsker m.fl. 2017; Srivastava, Li, och Dunson 2018). En fördel med uppdelning av data och beräkningar är att det lämpar sig väl för decentraliserad inlärning (Gürbüzbalaban m.fl. 2021) något som hamnat i alltmer fokus i och med ökad samhällsmedvetenhet kring hantering av data. För (2) så är det kritiska steget att avgöra vilket datapunkter som är redundanta och på sätt skapa en mindre datamängd som innehåller den essentiella informationen från den större mängden (Huggins, Campbell, och Broderick 2016). I och med att den mindre datamängden anses ha samma

egenskaper som hela datamängden så kan valfri algoritm sedan användas för träning (Campbell och Broderick 2019). Problemet med ansatsen är dock att den kräver en hel del steg och antaganden för att kunna användas och har därför omformulerats som ett VI-problem (Campbell och Beronov 2019). Vidare kan de olika stegen som datareduktion sedan träning vara ineffektivt och som följd av detta har det nyligen introducerats en metod som kombinerar de två i en helhet (N. Chen, Xu, och Campbell 2022).

3.2.3.2 Automation

En trend som belysts (Campbell och Broderick 2019) inom bayesisk dataanalys är den allt högre grad av beräkningsautomation som används framförallt i samband med träningsalgoritmer såsom t.ex. automatisk beräkning av gradienter för HMC (Carpenter m.fl. 2017), automatisering för att kalibrera hyperparametrar för träningsalgoritmer (M. D. Hoffman och Gelman 2014) samt automatik för att lösa optimeringsproblem utan att behöva härleda några direkta matematiska uttryck (Kucukelbir m.fl. 2017). Denna ökande grad av automatisering via mjukvara samt en alltmer större frihet vad gäller design och utformning av modeller via så kallad probabilistisk programmering (Carpenter m.fl. 2017), där modellen är frikopplad från träningsalgoritmen, gör att dataanalytiker har en större möjlighet att utforska olika typer av modeller och träningsansatser. Detta är vidare något som har gett upphov till att forskare inom området kopplat samman modelleringsprocesser med drag av processer inom mjukvaruutveckling (eng. software engineering), benämnt arbetsflöden (eng. workflows) (Gelman m.fl. 2020; Gabry m.fl. 2017). Stor vikt vid sådana flöden läggs på området modellkvalité där olika typer av tester kan göras för att utforska hur väl en modell fångat upp egenskaper från data, t.ex. (Gabry m.fl. 2017), eller hur väl en viss träningsalgoritm lyckats träna utifrån data. Kopplat till det finns metoder för att kalibrera (Talts m.fl. 2018), undersöka generalitet (Vehtari, Gelman, och Gabry 2017), samt olika verktyg för olika egenskaper av problem/data som kan användas (Yao, Yuling, Vehtari, Aki, och Gelman, Andrew 2022). Många utav dessa metoder har även publicerade mjukvarupaket kopplade till sig (Vehtari m.fl. 2022; Gabry och Mahr 2022).

3.2.3.3 Gaussiska processer

Gaussiska processer (eng. Gaussian processess) (GP) (Rasmussen 2004) är en typ av metod som benämns icke-parametrisk vilket syftar på att modellen inte har en statisk struktur med given dimensionalitet som direkt bestäms av ingående parametrar utan istället har en funktionell form som i sin tur bestäms av parameteriserade funktioner (av olika slag). Ordet icke-parametrisk kan här vara vilseledande då i själva verket parametrar ingår som en del av modellen. GP är ett högst aktivt område för forskning inom ML mycket delvis på grund av det finns intressanta samband med neuronnät och djupinlärning. Man kan nämligen visa att när vidden av en djupinlärningsmodell (baserat på neuroner) går mot oändligheten konvergerar modellen till en GP (vilket är att betrakta som en ”grund” modell) (Lee m.fl. 2017). Det finns också djupinlärningsvarianter av GPs där man bygger upp en modell hierarkiskt med en följd av GPs, så kallade djupa GP (Damianou och Lawrence 2013), men även dessa konvergerar mot en (grund) GP när bredden går mot oändligheten (Pleiss och Cunningham 2021) vilket har gjort att man vidare har studerat möjliga balanseffekter (eng. trade-off) mellan bredd och djup som då också relaterar till nätverksstruktur vid djupinlärningsmodeller. Med sådana intressanta sammankopplingar områdena emellan är det högst troligt att flera fortsatta studier kommande år kring relationen GP / djupa GP samt djupinlärningsmodeller kommer att utföras. Ett problem som kvarstår att lösa är att GPs är beräkningstunga så på liknande sätt som för generell skalbarhet inom bayesisk inferens så har man här inom detta område utfört mycket forskning för att kunna utföra inferens på större datamängder (H. Liu m.fl. 2020).

3.2.3.4 Övriga ansatser

En viktig komponent inom bayesisk modellering är att man på olika sätt kan nyttja priori fördelning som ett sätt att modellera olika önskvärda egenskaper i modellen. Ett exempel

på en sådan egenskap är så kallad ”gleshet” (eng. sparsity) som kan vara användbart när man har hög dimension vid t.ex. regression men där man tänker sig att det finns ett mindre antal faktorer som kan förklara målvariabeln och som då t.ex. gör det lättare för mänsklig tolkning. Här finns det olika specialtyper av priori fördelningar såsom ”spike-and-slab” samt hästsko (eng. Horseshoe) fördelning, se vidare (Piironen och Vehtari 2017) för en översikt. En spike-and-slab fördelning består av en mix av fördelningar där en av dem är huvudfördelningen (”slab”) som används för att modellera samband och den andra som är en mer koncentrerad fördelning kring noll (”spike”) och som används för att samla upp parametrar som har marginell eller ingen effekt på målvariabeln. Träning med en sådan fördelning gör att vissa koefficienter i regressionen tvingas mot noll, dvs. ingen effekt, och andra modelleras som i vanlig regression med huvudfördelningen. Ett annat sätt att modellera detta problem är med den redan nämnda hästskofördelningen vilken är uppbyggd kring att mer kontinuerligt förstärka vissa variabler och försvaga andra för att uppnå liknande effekt. Som tidigare nämnts kring bayesisk inferens så finns det skalbarhetproblem, så också i detta fall, vilket har lett till studier kring hur man kan utföra träning med dessa specialtyper av priors (Ray, Szabó, och Clara 2020; Johndrow, Orenstein, och Bhattacharya 2020).

En fördelning av annan karaktär, är den så kallade Dirichlet processen (Ferguson 1973; Antoniak 1974) vilken också tillhör kategorin inom bayesisk inferens benämnt icke parametrisk (eng. non-parametrics) där som tidigare beskrivits, i avsnitt 3.2.3.3, Gaussiska processer även ingår. Inom denna kategori återfinns modeller som i någon mening är dimensionslös, dvs. inte har en statisk struktur direkt kopplad till parametrar att estimera. En fördel med detta är att modellen inte är fastlåst till en given struktur före träning utan den lär sig strukturen under träningen. Dock så ökar komplexiteten med en sådan ansats och det kan vara svårt att träna en modell baserat på Dirichlet processen på ett effektivt sätt (Bryant och Sudderth 2012). Att använda denna typ av icke parametrisk fördelning har tidigare varit ett välstuderat området i synnerhet för olika typer av hierarkiska klusterproblem som använts inom textanalys (Teh m.fl. 2004) men nyligen har denna ansats även studerats tillsammans med djupinlärningsmodeller för att mer effektiv hitta en lämplig representation av kluster (N. Li m.fl. 2022).

3.3 Verktyg

Verktyg för dataanalys har haft en enorm framväxt och utveckling under en tid, sannolikt mycket tack vare att kommersiella aktörer (ofta benämnda ”techbolag”) verkat inom området speciellt vad gäller modellramverk för djupinlärning (Nguyen m.fl. 2019). Gemensamt för moderna verktyg inom dataanalysområdet är det utgör ett ramverk / ekosystem, många gånger med egna utvecklingskonferenser, som tillhandahåller olika funktioner samt moduler som är anpassningsbara. Detta har flera fördelar såsom att en dataanalytiker har stora möjligheter till att anpassa modellen och ändra relativt enkelt samt att modellstrukturen blir ganska transparant dokumenterad via ramverket jämfört med mer fristående kod skrivet i något mer generellt programmeringsspråk. Modellen är också särskild från träningsmetoden vilket gör att man relativt enkelt kan studera vilka effekter en träningsalgoritm har för en viss given modell. Ramverken har också ofta inom djupinlärning stöd för viss typ av specialhårdvara som kan utföra beräkningar effektivt. Vanligt förekommande ramverk inom djupinlärning är TensorFlow (Abadi m.fl. 2016) samt PyTorch (Paszke m.fl. 2019) där specialmoduler finns för att hantera fördelningar baseras på dessa ramverk såsom TensorFlow Probability[13] respektive Pyro (Bingham m.fl. 2019). Inom statistisk maskininlärning så är Stan (Carpenter m.fl. 2017) ett verktyg som omges av en stark ”forskningscommunity” inom bayesisk inferens.

[13] https://www.tensorflow.org/probability (besökt januari 2023)

3.4 Framtidsprognos och analys

I takt med en allt mer utbredd användning av AI, där dataanalys ingår som en central del, har krav på AI-metoder blivit ett område med ett allt större fokus. Detta manifesterar sig på olika sätt bland annat genom förslag för reglering av AI på EU-nivå men även i olika typer av nya forskningsområden som belyser önskvärda krav på AI såsom förklarbarhet (eng. explainability) (Arrieta m.fl. 2020). Vidare så har man på EU-nivå sammanställt en kriterielista för tillförlitlig AI (eng. trustworthy AI) där ett utav områdena som osäkerhetsmodellering skulle falla inom är ”teknisk robusthet och säkerhet” (eng. ”technical robustness and safety”).

Förutom denna uppmärksamhet kring olika önskvärda egenskaper hos AI-metoder på EU-nivå så har även många tillämpningsområden där AI används i olika utsträckning någon form av riskkomponent. Vidare finns det många system / miljöer som är icke- deterministiska och där således olika former av osäkerheter existerar naturligt vilket då också ofta ställer krav på förmåga att representera osäkerhet hos AI-metoder.

Vikten och aktualiteten av området belyses också genom de flertal områdesartiklar som nyligen givits ut och som refererats i denna rapport. Med tanke på den alltmer utbredda användningen och ökade medvetenheten om AI-metoder så är det högst sannolikt att än mer krav kommer att ställas även inom områden som inte direkt innehar en riskkomponent, t.ex. genom implicita krav från konsument eller där konkurrens gör att mindre lämpade modeller slås ut. Om man ser prediktioner som uttalande från AI så är det inte mer än rimligt att representera nyanser av dessa i termer av osäkerheter speciellt med tanke på att många sådana prediktioner är framtidsuttalanden vilket av naturliga skäl innehar osäkerhet. Avgörande för i vilken utsträckning dataanalysmetodik med osäkerhetsmodellering kommer att användas är dock sannolikt vad man väljer att göra med denna osäkerhet både från ett perspektiv av mänskligt beslutsfattande men även i mer automatiserade system.

Vad gäller de två olika huvudområdena listade i denna rapport, dvs. djupinlärning och statistisk maskininlärning, så har de som tidigare nämnts i och med det ökade fokuset på bayesisk djupinlärning alltmer växt ihop utifrån synvinkeln att resultat från ett område överförts till det andra, framförallt från statistisk maskininlärning till djupinlärning i och med det ökade intresset av osäkerhetsmodellering inom det området. Sannolikt kommer så även ske i framtiden där då djupinlärning kanske kommer att ses som en viss typ av modellstruktur som då kräver en viss typ av träningsmetod för att erhålla önskvärda resultat i termer av osäkermodellering. Vidare så kan det finnas goda möjligheter för statistisk maskininlärning att vidare utvecklas baserat på resultat inom djupinlärning, kanske framförallt med avseende på olika tekniker för att erhålla mer skalbara lösningar, något som denna rapport belyst utgöra en kärnproblematik och som är ett aktivt forskningsområde. Fortsatt utredning av vilken design av djupinlärningsmodell, t.ex. i termer av antal lager och neuroner i relation till val av träningsalgoritm för att erhålla prediktioner av en viss given kvalité kommer sannolikt vara avgörande för användning inom mer tillämpningsorienterade områden.

4 Förklarbarhet

2017 lanserade *Defense Advanced Research Projects Agency* (Darpa) projektet *eXplainable AI* (XAI) vilket översatt benämns *förklarbar AI*. Projektet motiverades med de lovande framsteg inom *maskininlärning* (eng. *machine learning*, ML) som medfört nya utmaningar. I takt med ökad prestanda har komplexiteten skenat varpå ML-modellerna upplevs som icke-transparenta, dvs. utan insyn i hur de drar sina slutsatser vilket minskar tilltron till modellerna. Ett vanligt tillkortakommande för AI-systemen är oförmågan att förklara beslut och handlingar för den mänskliga användaren. XAI utgör en grund i arbetet att förstå och skapa förtroende för AI-system genom särskilt utvecklad metodik, ofta med hänsyn till psykologiska mekanismer (Gunning och Aha 2019). Även EU:s AI act[14] lyfter fram XAI som en viktig teknik för att kommande lagstiftning skall kunna införas.

Modellprestanda och förklarbarhet beskrivs ofta som ett kompromissförhållande där en högre grad av den ena sker på bekostnad av den andra. Denna motsats ämnar XAI att överbrygga eftersom de två aspekterna behöver samexistera i ett tillförlitligt system. Däremot är processen att mäta ett systems förklarbarhetsgrad aningen komplicerad eftersom det för närvarande vanligen kräver mänsklig utvärdering, lämpligen genom psykologiska experiment, och därmed inte kan kvantifieras internt likt prestandautvärdering (Gunning och Aha 2019).

Darpas XAI-projekt konkluderade vid avslut år 2021 att det för närvarande inte finns en universell lösning för XAI, utan att olika typer av AI-metoder och användare kräver olika förklaringsverktyg. En av de stora utmaningarna är hur man mäter effektivitetsgraden av en förklaring, men projektet fastslår att användare generellt föredrar system med förklaringar jämfört med förklaringslösa beslut. Särskilt användbart kan stödet av förklarbarhet vara vid prediktering med indata som är främmande för modellen (i syfte att bedöma tillförlitligheten för utmatningen) eller som stöd i att härleda orsakerna till felaktiga beslut tagna av AI-system (Gunning m.fl. 2021).

Darpas projekt är emellertid inte nyskapande, utan bör främst betraktas som en framgång genom att fokus återigen riktas mot dessa aspekter – delvis genom själva namnsättningen – efter ett långvarigt fokus inom fältet på AI-modellers prediktiva förmåga. Redan i mitten av 1970-talet studerades förklaringar av expertsystem, och sedan millennieskiftet även av moderna AI-metoder som *neuronnät* (eng. *neural networks*). XAI bör inte betraktas som renodlad teknik, utan snarare som ett samlingsbegrepp representerandes initiativ för att åstadkomma egenskaper så som transparens och begriplighet. Således förekommer alternativa benämningar med delvis andra aspekter, däribland *tolkningsbar ML* (eng. *Interpretable ML*), och *ansvarsfull AI* (eng. *Responsible AI*) (Adadi och Berrada 2018; Carvalho, Pereira, och Cardoso 2019).

Kapitlet är strukturerat enligt att först med avsnitt 4.1 introducera en taxonomi som tillämpas. I avsnitt 4.2 beskrivs aspekten av utvärdering och jämförelser. Sedan redogörs i avsnitt 4.3 för en generell lägesbild för utvecklingen. Vidare kartläggs i avsnitt 4.4 tillämpningar för olika observerade delområden. Senare listas i avsnitt 4.5 status och utsikter för militära tillämpningar. Avslutningsvis presenteras i avsnitt 4.6 slutsatser baserat på de föregående avsnitten.

4.1 Taxonomi

Ett genomgående hinder för XAI-fältet är bristen på en vedertagen entydig taxonomi, vilket flertalet översiktsartiklar, exempelvis Adadi and Berrada (2018), Barredo Arrieta et al. (2020) och Speith (2022) har konstaterat. Diverse återkommande nomenklatur används, men emellertid varierar betydelsen och viss litteratur tillämpar begreppen synonymt. Vad

[14] https://artificialintelligenceact.eu/ (besökt augusti 2023)

som bedöms vara en användbar och återkommande precisering av central nomenklatur lyder (Barredo Arrieta m.fl. 2020):

- *Tolkningsbarhet* (eng. *interpretability*): passiv egenskap som en AI-metod eller modell besitter som gör det möjligt för människan att förstå den.
    - *Transparens* (eng. *transparency*): används helt synonymt med tolkningsbarhet.

- *Förklarbarhet* (eng. *explainability*): aktiv egenskap som inbegriper handlingar för att förtydliga ML-modellens funktion för människan.

En frekvent förekommande särskiljning är den mellan *transparenta*[15] modeller vilka beskrivs besitta en inneboende förklarbarhet benämnt *ante-hoc*, och *black-box* modeller med hänvisning till dess icke-transparenta systemfunktion som behöver ett särskilt lager för att uppnå förklarbarhet, benämnt *post-hoc*. Emellertid är uppdelningen delvis problematisk då en metod som anses vara ante-hoc i grunden kan potentiellt förlora den egenskapen vid ökad komplexitet genom exempelvis fler parametrar eller beslutsregler (Barredo Arrieta m.fl. 2020; Speith 2022).

Ante-hoc kan betraktas på en hierarkisk nivå, vilken presenteras i fallande ordning där en lägre nivå utgör en delmängd av en högre nivå (Barredo Arrieta m.fl. 2020):

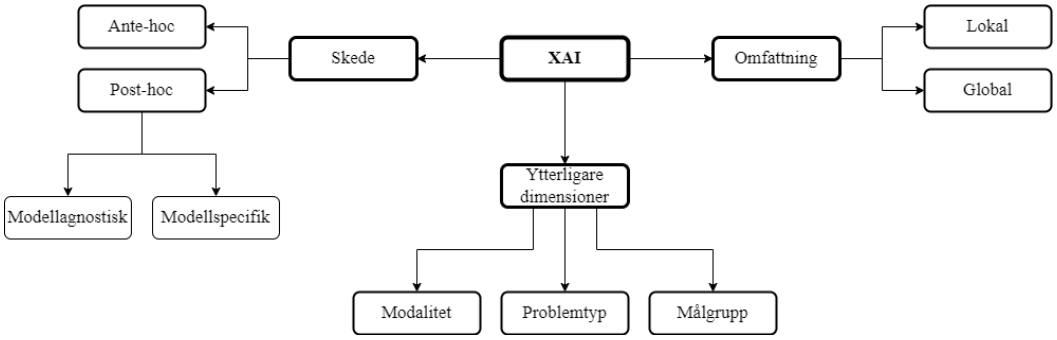
1. *Simuleringsbarhet* (eng. *simulatability*) vilket inbegriper att en modell i sitt utförande är tillräcklig för att en människa ska kunna tänka och resonera om den i sin helhet.
2. *Nedbrytbarhet* (eng. *decomposability*) vilket inbegriper förmågan att förklara alla delar av en modell, t.ex. inmatning, parametrar och beräkningar.
3. *Algoritmisk transparens* (eng. *algorithmic transparency*) vilket inbegriper förmågan att oavsett inmatning är det möjligt att förutspå modellens utmatning.

Post-hoc metoder delas framförallt upp i *modellspecifika* och *modelloberoende* metoder, bestående av bland annat följande övergripande tekniker (Adadi och Berrada 2018; Speith 2022):

- *Visualiseringar* som genom grafiska metoder ger insikter i modellens funktionalitet.
- *Kunskapsextrahering* från interna processer för modellen som ger ökad förståelse.
- *Påverkansmetoder* vilka estimerar effekten på utmatningen genom att ändra inmatning eller interna komponenter.
- *Exempelbaserat* som förklarar modellens utmatning genom selektering av vissa observationer i datamängden.

En ytterligare frekvent gruppering av metoder består av uppdelningen i form av huvudsakligen *lokala* och *globala* förklarbarhetsmetoder. Med ett lokalt perspektiv avses förklaringar av individuella observationer, medan ett globalt perspektiv ämnar beskriva modellens fulla beteende (Adadi och Berrada 2018). Med inspiration av Speith (2022) presenteras i figur 6 ett möjligt schema över taxonomin. Ytterligare dimensioner att betrakta inkluderar exempelvis modalitet (bland annat bilder eller text), problemtyp (bland annat regression eller klassificering) och målgrupp (bland annat utvecklare, beslutsfattare eller slutanvändare).

[15] Alternativa förekommande benämningar är *white-box* eller *glass-box*, som kontrast till *black-box*.



Figur 6: Taxonomischema för XAI.

4.2 Utvärdering och jämförelse

Utöver utmaningen för en metod att tillhandahålla förklarbarhet som tillför värde för den mänskliga mottagaren, omfattas även en problematik gällande hur rättvisande och tillförlitliga förklaringarna är, vilket främst berör post-hoc metoder. Önskvärda egenskaper har presenterats i diverse litteratur, men hindras av hur de praktiskt kan kvantifieras eller enhetligt ska mätas. Egenskaper som önskas representeras med metriker är bland annat (Carvalho, Pereira, och Cardoso 2019):

- *Begriplighet* (eng. *comprehensibility*) vilket motsvarar hur väl den mänskliga användaren kan förstå och tillägna sig förklaringen.
- *Noggrannhet* (eng. *fidelity*) vilket motsvarar hur väl förklaringsmetoden approximerar modellen av intresse. I viss litteratur används begreppet *träffsäkerhet* (eng. *accuracy*) men då riskeras det att förväxlas med det traditionella prestandamåttet för klassificering.
- *Stabilitet* (eng. *stability*) vilket motsvarar hur överensstämmande förklaringar för likartade indata är för en given modell.
- *Konsekvens* (eng. *consistency*) vilket motsvarar hur pass lika förklaringar är som erhålls från två olika modeller vilka är tränade för samma syfte med samma data.

I en kartläggning av XAI-metoder konstaterar Confalonieri m.fl. (2021) att de flesta post-hoc metoder saknar förmågan att generera garantier avseende osäkerheten i förhållande till den underliggande modellens funktionalitet. Samtidigt konstaterar Ras m.fl. (2022) ett ökat forskningsengagemang för utvärderingsmetoder för XAI-metoder som en konsekvens av att vissa tillgängliga metoder konstaterats generera felaktiga förklaringar.

XAI baseras på ett samspel mellan maskin och människa, varför prestandautvärderingen troligen inte går att isolera helt från utvärdering även av den mänskliga användarens prestation. Tillförlitligheten kan delvis bero på uppnådda resultat jämfört med en användare utan det undersökta stödet. Utöver aspekter som responstid och nivå av korrekthet, är sannolikt en viktig faktor vilka fel som system begår jämfört med fel som människor begår (R. R. Hoffman m.fl. 2019).

4.3 Generellt forskningsläge

Behovet av XAI går att betrakta ur fyra huvudsakliga förklarbarhetsperspektiv (Adadi och Berrada 2018):

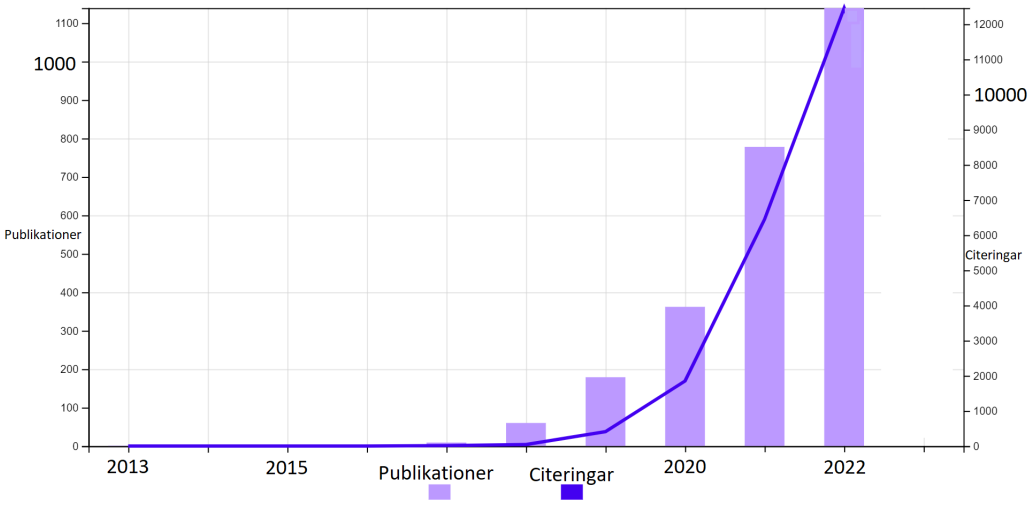
- *Rättfärdiga* varför AI-systemet ger upphov till ett visst resultat för att exempelvis motsvara krav som beslutsfattande utan bias.
- *Kontrollera* att AI-systemet är korrekt konstruerat och förstå dess begränsningar.
- *Förbättra* prestandan för AI-systemet genom insikter av dess funktionalitet i en kontinuerlig iterativ process.

- *Upptäcka* ny kunskap som AI-systemet lärt sig som tidigare var okänt för människan, exempelvis nya strategier i spel.

Inriktad XAI-forskning inbegriper främst sektorer där kostnaden av felaktiga prediktioner är påtaglig, bland annat transport med exempelvis självkörande fordon, sjukvård med exempelvis automatiserad diagnostisering, juridik med exempelvis bedömning av återfallsrisk, och finans med exempelvis system för kreditbedömningar (Adadi och Berrada 2018). FOI har tidigare kartlagt XAI för *djupinlärning* (eng. *deep learning*) inriktat på militära tillämpningar (Luotsinen m.fl. 2019).

En omfattande kartläggning är genomförd av Linardatos, Papastefanopoulos, och Kotsiantis (2020) där föreslagna metoder presenteras med antal citeringar per år och en enhetlig klassificering av metoderna samt dess tillämpningsmodalitet. Bland studiens slutsatser konstateras att XAI för djupinlärning har ägnats stor uppmärksamhet för forskning, och särskilt tillämpningar för *datorseende* (eng. *computer vision*) och *språkteknologi* (eng. *natural language processing*).

Resultat från sökverktyget *Web of Science*[16] avseende utvecklingen av XAI över tid presenteras i figur 7 och forskningens ursprungsländer syns i figur 8. Antalet publiceringar presenteras på den vänstra vertikalaxeln med tillhörande stödlinjer, medan den högra vertikalaxeln motsvarar antalet citeringar. Det finns inga tecken på en avmattning för området utan antalet publiceringar och citeringar påvisar ett ämne med god utveckling. Den iakttagna trenden stöds av (Linardatos, Papastefanopoulos, och Kotsiantis (2020) som noterar en snabb tillväxt, men framhåller bristen på enighet inom området som ett hinder. Mängden studier tolkas som ett bevis på områdets nytta, samtidigt bedöms området besitta outforskade aspekter med potential att upptäckas inom de närmaste åren.



Figur 7: Sökning i Web of Science den 30 juni 2023 genom *TS=("Explainable AI" or "XAI")* med erhållet H-index[17] 55.

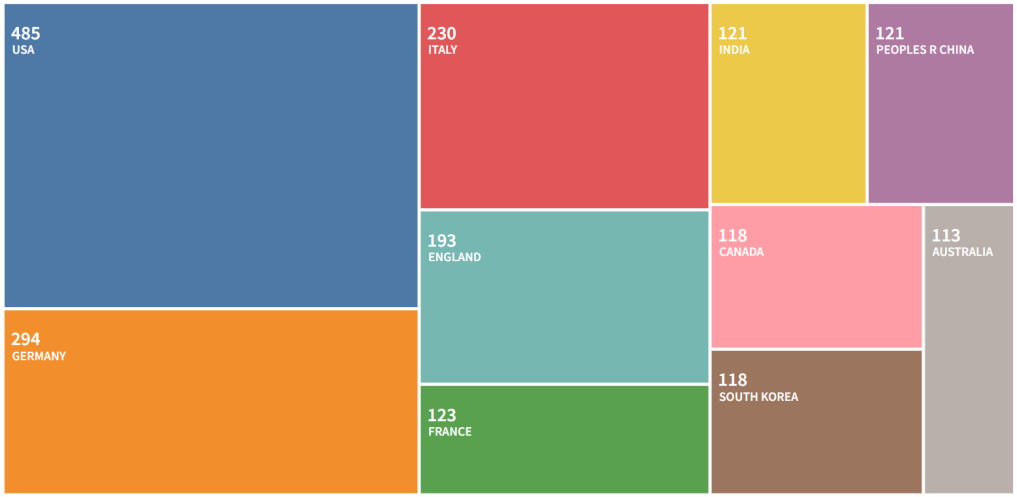
Ytterligare indikationer på utvecklingstrenden består exempelvis av EU:s (Europeiska unionens) strategi för AI vilken är inriktad på spetskompetens och förtroende (Europeiska kommissionen 2023), med koppling till *right to explanation*[18]. Ett annat exempel utgörs av

[16] WoS-sökfråga: (TS=(XAI or (explainable AI)) AND (SU = (Computer Science) OR SU = (Mathematics) OR SU = (Automation & Control Systems) OR SU = (Engineering) OR SU = (Robotics) OR SU = (Science & Technology Other Topics) OR SU = (Telecommunications))) AND PY=(2012-2022)

[17] https://support.clarivate.com/ScientificandAcademicResearch/s/article/Web-of-Science-h-index-information?language=en_US (besökt mars 2023)

[18] https://en.wikipedia.org/wiki/Right_to_explanation (besökt mars 2023)

OpenAI[19] som vid lansering av GPT-4[20] presenterade sin syn på framtiden där en del kretsade kring tolkningsbarhet och förklarbarhet för att adressera de begränsningar som medförs av nuvarande black-box AI-metodik (OpenAI 2023).



Figur 8. Forskningens ursprungsländer för söktermen ”XAI” or ”explainable AI” år 2013–2022

Flera konferenser har haft vad som benämns som *workshops* eller *special sessions* inriktade på XAI. Aktuella konferenser är exempelvis IEEE:s[21] konferens IJCNN[22] för 2023, eller ACM:s[23] konferens CHI[24] med HCXAI[25]. Vidare finns det numera konferenser helt ägnade åt XAI-aspekter, däribland ACM FaacT[26] och XAI 2023[27].

Som stöd för implementering av XAI-metoder finns flera mjukvaror tillgängliga. Vanligen utgörs det av bibliotek för Python[28] med vissa resurser i utvecklingsstarten medan andra består av stabila versioner. Några nämnvärda bibliotek med samlingar av algoritmer är Captum[29] (tillhandahållet av Meta[30]) och InterpretML[31] (utvecklat av Microsoft[32]). Vissa algoritmer har även egen tillhandahållen mjukvara. Kartläggningen av Dwivedi m.fl. (2023) rekommenderas för mer ingående information gällande tillgänglig mjukvara relaterad XAI.

4.4 Kartläggning per delområde

Utvecklingen inom AI sker som bekant i en imponerande takt. Som beskrivet i avsnitt 4.2 är en utmaning för delområdet XAI hur metoder skall jämföras och utvärderas. Således är

[19] https://openai.com/ (besökt mars 2023)

[20] https://openai.com/product/gpt-4 (besökt mars 2023)

[21] https://www.ieee.org/ (besökt mars 2023)

[22] https://2023.ijcnn.org/ (besökt mars 2023)

[23] https://www.acm.org/ (besökt mars 2023)

[24] https://dl.acm.org/conference/chi (besökt mars 2023)

[25] https://hcxai.jimdosite.com/ (besökt mars 2023)

[26] https://facctconference.org/ (besökt mars 2023)

[27] https://xaiworldconference.com/ (besökt mars 2023)

[28] https://www.python.org/ (besökt mars 2023)

[29] https://captum.ai/ (besökt mars 2023)

[30] https://opensource.fb.com/ (besökt mars 2023)

[31] https://interpret.ml/ (besökt mars 2023)

[32] https://www.microsoft.com/ (besökt mars 2023)

det en minst sagt utmanande uppgift att identifiera den ledande tekniken. Därför baseras följande redovisning främst på beskrivande kategorisering kring vad som kan liknas vid metodfamiljer. Selekteringen av de algoritmer som redovisas baseras till hög grad på popularitetsvärdering gällande antalet tillämpningar i litteraturen i kombination med aktualitet.

En ej uttömmande kartläggning redovisas genom först avsnitt 4.4.1 som avhandlar modelloberoende metoder. I avsnitt 4.4.2 och 4.4.3 avhandlas uppgifterna datorseende respektive språkteknologi. Sedan redogörs i avsnitt 4.4.4 för särskilda riktlinjer för problemuppgiften regression. Vidare presenteras i avsnitt 4.4.5 multimodalt lärande. Senare kartläggs förstärkningsinlärning i avsnitt 4.4.6. Slutligen listas i avsnitt 4.4.7 några övriga aspekter som knyter an till transparenta modeller.

4.4.1 Modelloberoende metoder

Modelloberoende (eng. model agnostic) förklaringsmetoder är tillämpbara oberoende av vilken modell som betraktas varför metoderna kategoriseras som post-hoc. Följaktligen kan de tillämpas i diverse sammanhang och återanvändas oberoende av modellen i fråga. Emellertid kan modellspecifika förklaringsmetoder anses vara mer användbara och informativa vid betraktande av specifika fall (Carvalho, Pereira, och Cardoso 2019).

Bland de modelloberoende förklaringsmetoderna för att förklara black-box modeller dominerar metoderna *LIME* (*Local Interpretable Model-agnostic Explanations*) (Ribeiro, Singh, och Guestrin 2016) och *SHAP* (*SHapley Additive exPlanations*) (Lundberg och Lee 2017) användningen i litteraturen enligt Linardatos, Papastefanopoulos, och Kotsiantis (2020). LIME baseras på surrogatmodeller som lokalt approximerar modellen som önskas förklaras i syfte att bidra med en kvalitativ förklaring mellan inmatning och utmatning (Ribeiro, Singh, och Guestrin 2016). Med ursprung i den spelteoretiska metoden *Shapley value*[33] presenteras SHAP som en teoretiskt välgrundad metod som förenar diverse separata föreslagna metoder som baseras på additiva särdragsförklaringar (Lundberg och Lee 2017). Gemensamma drag för LIME och SHAP är egenskaperna att kvantifiera och visualisera betydelsen av särdrag (eng. feature) för den betraktade modellens prediktioner. Genom analys av särdragen undersöks modellens uppskattade funktionalitet och därav möjliggörs ökad förklarbarhet (Linardatos, Papastefanopoulos, och Kotsiantis 2020).

Kritik har riktats mot såväl LIME som SHAP, exempelvis av Främling et al. (2021) som hävdar att de metoderna uppskattar påverkan (eng. influence) som en kombination av betydelse (eng. importance) och nytta (eng. utility). Som ett alternativ har *CIU* (*Contextual Importance and Utility*) (Främling 2020) presenterats och kan tillämpas för icke-linjära modeller vilket i vissa fall är en begränsning för andra metoder. Noterbart är att CIU har utvecklats från beslutsteori (eng. decision theory), en domän som enligt Främling (2020) är underexploaterad i XAI-kontexten trots lång historik och hög relevans.

4.4.2 Datorseende

En kartläggning av lokala post-hoc förklaringsmetoder för klassificering i datorseende med neuronnät (exempelvis faltningsbaserade, eng. convolutional, CNN) är genomförd av Nielsen m.fl. (2022). En uppdelning i metoder baseras på:

- *Särdragsstörning* (eng. *feature perturbation*) som baseras på förändring av särdragen för att observera effekten av handlingen på modellprestandan. Metodiken kräver flera utvärderingar av modellen för att utvärdera pixlarna i bilden varför det är beräkningsintensivt.

[33] https://en.wikipedia.org/wiki/Shapley_value (besökt mars 2023)

- *Gradientinformation* (eng. *gradient information*) som använder gradienten för utmatning givet inmatning för att uppskatta effekten av särdragen på modellprestandan. Gradienterna är vanligen brusiga varför metodiken kan resultera i felaktiga förklaringar.

Kartläggningen av Nielsen m.fl. (2022) avgränsas vidare till metoder baserade på gradientinformation, specifikt *tillskrivningskartor*[34] (eng. *attribution maps*) vilka graderar betydelsen för den spatiala informationen för klassificeraren gällande den betraktade inmatningen för att generera ett nytt abstraktionslager. Eftersom metoderna är lokala och betraktar individuella observationer uppstår en eventuell problematik om modellen i helhet försöks förklaras. Vidare medför tillämpningen av post-hoc metoder två eventuella felkällor: de relaterade till modellen och de introducerade av förklaringsmetoden. Facit (eng. ground truth) saknas för förklaringarna, och följaktligen är det en komplicerad uppgift att särskilja felkällorna vilket bidrar med ökad osäkerhet (Nielsen m.fl. 2022).

Enligt Nielsen m.fl. (2022) råder ingen konsensus kring vilken förklaringsmetod som är att föredra, vilket delvis kan härledas till avsaknaden av metriker för att rättvist jämföra metoderna. Däremot kan förekomst av metoders inneboende förmågor till viss del jämföras. Ett exempel är förmågan att förklara sambandet mellan inmatning och utmatning för en betraktad modell. Det kräver metodik som beaktar utmatningsklass, vilket vissa föreslagna metoder inte inbegriper, så kallat invariant-klass-beteende (Nielsen m.fl. 2022).

En metod med dominerande antal citeringar är *Grad-CAM* föreslagen av Selvaraju m.fl. (2017) som således används frekvent. Metoden baseras på nyttjande av det sista faltningslagret i modellen för att producera en karta över regioner som är betydelsefulla för prediktionen för den betraktade klassen givet inmatningen (Linardatos, Papastefanopoulos, och Kotsiantis 2020). Dessa typer av metoder är användbara för att kontrollera att besluten baseras på lämplig del av informationen, exempelvis för optiska bilder så är det vanligen objekten i fråga och inte bakgrunden.

Den andra punkten i uppdelningen av metoder, särdragsstörning, har ägnats viss uppmärksamhet av Ras m.fl. (2022) i en omfattande kartläggning av XAI för djupinlärning. Bland de metoder som listas baseras de vanligen på systematisk maskning av pixlar för att undersöka påverkan på modellens prediktioner, eller genom att tillföra brus eller oskärpa. Metoderna beskrivs vanligen som användbara för känslighetsanalys genom markeringar av regioners vikt för klassificeringen vilket användaren kan använda för att förstå vilka delar av bilden som bidragit till klassificeringen (Ras m.fl. 2022).

4.4.3 Språkteknologi

De senaste framstegen för språkteknologimodeller har baserats på neuronnät vilket återspeglas i föreslagna XAI-metoder. Bland post-hoc metoder är följande grupperingar vanligt använda (Danilevsky m.fl. 2020):

- Gradientbaserade metoder, särskilt de som inriktas på tillskrivningar genom första derivatan vilket kan tillämpas för att beräkna särdragsbetydelse på ordnivå.
- *Layer-wise relevance propagation* (Bach m.fl. 2015) ursprungligen föreslaget för bilder men adapterats till språkdomänen bland andra. Metoden baseras på att beskriva relevansen för särdrag i ett valt neuronlager genom sekventiell tillämpning genom bakåtpropagering (eng. backpropagation).

Metoderna kombineras ofta med visualiseringstekniker för att åskådliggöra vilken tillgänglig information som används i modellen. Föreslagna XAI-metoder lanseras däremot vanligen utan väl underbyggd motivering för dess lämplighet i att bidra med förklarbarhet.

[34]För tillhörande visualiseringar förekommer benämningarna *saliency maps* och *heatmaps* (Ras m.fl. 2022).

Det är troligen som följd av avsaknaden av konsensus för hur förklaringsbarheten skall utvärderas (Danilevsky m.fl. 2020).

Uppmärksamhetsbaserad djupinlärning för språkteknologimodeller har historiskt betraktats som självförklarande. På senare tid har det till viss del konstaterats tvivelaktigt eftersom sambandet mellan visualiseringar av uppmärksamhetsmarkeringar och modellens utmatning vanligen är komplext, och i förekommande fall även svagt (Ras m.fl. 2022).

Enligt Cambria m.fl. (2023) är det tidskrävande att välja lämplig XAI-metod och riskerar att bli felbenäget som en konsekvens av bristande nyttjande av lämpliga presentationstekniker.

Som målbild förespråkas ett verktyg i vilket användare kan konversera med AI-systemet för att gemensamt resonera kring beslutsfattande likt processen mellan människor. Vidare observeras det en snabb tillväxt för antalet föreslagna metoder, men även en ökande grad av metoders relevans och nytta.

4.4.4 Regression

Enligt Letzgus m.fl. (2022) är den absoluta majoriteten av XAI-forskningen inriktad på klassificeringsmodeller. Som en konsekvens är tillgången på XAI-metoder särskilt utvecklade för regressionsmodeller begränsad varför XAI-metoder utvecklade för klassificering vanligen appliceras även för regression. Resultaten kan uppvisa nytta, men emellertid kan de två uppgifterna – klassificering och regression – beskrivas bestå av fundamentala konceptuella skillnader varför särskild hänsyn lämpligen bör vidtas (Letzgus m.fl. 2022).

Vissa XAI-metoder för klassificering bidrar med klasspecifika förklaringar varför beslut om varför den ena klassen framför en annan kan jämföras. För regressionsfallet rekommenderas referensvärden användas vilket möjliggör liknande tillvägagångssätt, men lösningen kräver att metoder har utvecklats för användning av referensvärden, vilket är sällsynt förekommande (Letzgus m.fl. 2022).

I redogörelsen av Letzgus m.fl. (2022) för tillgängliga post-hoc metoder rekommenderas förklaringsmetoder som baseras på en bevarandeprincip (exempelvis Shapley values) eftersom det tillåter uppdelning av beståndsdelarna för modellens prediktion för utmatningen. Noterbart är att för regressionsmodeller bedöms utmatningen potentiellt bestå av mer information än för klassificeringsfallet. Således är det fördelaktigt i förklarbarhetssynpunkt att tillämpa en mätskala för utmatningen som är naturligt representativ för uppgiften som undersöks, exempelvis fysikaliska måttenheter eller monetära enheter, i kontrast till transformationer för ett modelleringssyfte (Letzgus m.fl. 2022).

4.4.5 Multimodalitet

Multimodala AI-modeller har uppvisat imponerande resultat de senaste åren, även för unimodala tillämpningar. Det är vanligt att fusionera datorseende och språkteknologi med tillämpning av neuronnät. Multimodal AI medför flera komplexa aspekter däribland hur modaliteterna skall representeras, översättas, och fusioneras. Den mänskliga perceptionen är multimodal varför förklaringar av multimodala AI-system är centralt i arbetet att konstruera intelligenta, resonerande system som kan interagera förtjänstfullt med människor (Joshi, Walambe, och Kotecha 2021).

Representation genom olika modaliteter har bistått med ökad prestanda. Men de multimodala systemen kan även dra fördel av modaliteternas olika styrkor avseende förklaringsmöjligheter varför ytterligare synergieffekter kan uppnås. Emellertid skall det nämnas att ett multimodalt system i grunden är mer komplext att förstå (Ras m.fl. 2022; Joshi, Walambe, och Kotecha 2021). Framstegen med multimodal XAI är begränsade och består främst av ett konceptuellt resonemang snarare än demonstration med framgångsrika metoder, om än utveckling pågår. Ett centralt hinder är utvärderingsprocessen av genererade förklaringar och säkerställning att de är modellen trogna (Joshi, Walambe, och Kotecha 2021).

En föreslagen lösning är tillämpningen av neuronnät med grafrepresentation av data (eng. graph neural network[35], GNN) som utöver imponerande prestanda även förespråkas representera samband som är mer förståeliga för mänskliga användare. GNN:er är dock vanligen att betrakta som black-box modeller eftersom be består av komplexa samband med stort omfång. Som en lösning har post-hoc metoder som *GNNExplainer* (Ying m.fl. 2019) föreslagits i syfte att generera förklarbarhet för prediktioner genom att nyttja de relationer som modellerats i grafrepresentationen och identifiera den delmängd som består av mest relevant information (Holzinger m.fl. 2021).

Konceptet med multimodalt lärande är besläktat med gemensamt lärande (eng. multi-task learning, alternativt eng. joint training) där flera uppgifter löses samtidigt. För XAI kan systemet konstrueras enligt att en sekundär uppgift löses vid sidan av huvuduppgiften som ämnas lösas, där sekundäruppgiften renderar direkta eller indirekta förklaringar för huvuduppgiftens beslutsfattande (Ras m.fl. 2022).

4.4.6 Förstärkningsinlärning

Konceptuellt skiljer sig *förstärkningsinlärning* (eng. *reinforcement learning*, *RL*) från övrig maskininlärning som beskrivs som *övervakad* (eng. *supervised*) eller *oövervakad* (eng. *unsupervised*). Förstärkningsinlärning är på inget sätt enklare, men lärandeformen besitter andra möjligheter (och begränsningar) varför särskilda metoder krävs. Centrala aspekter att förklara är exempelvis en policy[36] eller varför en agent tog ett visst beslut i ett visst läge (Heuillet, Couthouis, och Díaz-Rodríguez 2021).

I en kartläggning av agentbaserad förstärkningsinlärning av Heuillet, Couthouis, och Díaz-Rodríguez (2021) konstateras att ingen universalmetod för att generera förklarbarhet existerar. Tillgängliga XAI-metoder (i sammanhanget även benämnt *XRL*) är i allmänhet modellspecifika samt tillämpningsberoende, med utgångspunkt vanligen från robotik eller spel. Metoder baserade på Shapley values har föreslagits som en möjlig väg framåt för att generera mer generella verktyg.

Avseende tillgängliga XAI-metoder beskrivs de som bistår utvecklare som mest lovande för närvarande. Det gäller främst metoder som baseras på en hierarkisk struktur där en agent på en hög nivå bryter ner uppgiften i deluppgifter vilka fördelas till andra agenter på en lägre nivå (Heuillet, Couthouis, och Díaz-Rodríguez 2021). Ett exempel på en sådan hierarkisk metod är *Hindsight Experience Replay* (Andrychowicz m.fl. 2017) som möjliggör lärande genom avklarande av deluppgifter som tillsammans utgör helhetslösningen, men i en representation vilken är mer förståbar.

4.4.7 Övrigt

Vad som är att betrakta som transparenta metoder beror till stor del på avsedd målgrupp. Däremot finns det konsensus kring att vissa metoder i grunden är mer förklaringsbara än andra. Flera XAI-metoder som fått genomslag baseras på post-hoc, samtidigt som kritik har riktats mot förhållningssättet vilket kan beskrivas som en tvåstegslösning. Bland annat argumenterar Rudin (2019) för att lösningen måste bestå i att från grunden utveckla metoder som är förklaringsbara för att säkerställa att förklaringarna är trogna modellens funktionalitet.

Ett koncept som relaterar till transparenta modeller är *modelldestillering* (eng. *model distillation*) som består av tekniker för att reducera komplexiteten för ett (djupt) neuralt nätverk till en representation som är förståbar för användaren. Den destillerade modellen behöver inte nödvändigtvis underprestera jämfört med originalmodellen eftersom insikter kan överföras. Metodiken har beröringspunkter med lokala surrogatmodeller (se avsnitt

[35] https://en.wikipedia.org/wiki/Graph_neural_network (besökt mars 2023)

[36] En policy är en inlärd funktion som föreslår en handling för alla tänkbara situationer.

4.4.1), men vid framställning av en global surrogatmodell förekommer vanligen benämningen *modellöversättning* (eng. *model translation*). Användningen består i att ersätta originalmodellen eller alternativt som ett verktyg för att formulera hypoteser avseende utmatningen för originalmodellen (Ras m.fl. 2022).

4.5 Militär tillämpning

I kontexten av autonoma vapensystem i synnerhet och militär tillämpning av AI i allmänhet avhandlar Holland Michel (Holland Michel 2020) förklarbarhet som en komponent i vad som benämns som *förståbarhet* (eng. *understandability*). Det används som förtydligande av vad som beskrivs som inkluderande av mänskliga aspekter utöver tekniskt orienterade aspekter. Det problematiseras vad som räknas som förståbarhet (alternativt förklarbarhet) och vad för kravställning som bör gälla utefter diverse tillämpningsfall. Autonoma vapensystem kringgår inte ansvarsutkrävande varför god tillförlitlighet och förutsägelsebarhet är nödvändigt. Därför värderas området högt för framgångsrika AI-system men betraktas än så länge fortfarande vara i begynnelsen.

En annan bedömning av teknologiska utsikter för XAI är genomförd av Hult m.fl. (2022) som bedömer områdets potential besitta signifikant militär nytta. Det förutspås större tillgängliga datamängder från ökad tillämpning av sensornätverk varför analysmetoder ökar i betydelse eftersom innehållet blir för omfattande för enbart mänsklig hantering. XAI spås få ökad betydelse på stridsteknisk, taktisk och operativ ledningsnivå för att genomföra analys och bygga tillit. Konkreta tillämpningar inbegriper troligen exempelvis cyberattacker, spaning och medicinsk behandling. Bedömningen förefaller dock vara att XAI fortfarande är i ett tidigt utvecklingsstadium för denna tillämpning.

En konkret militär tillämpning redovisas av Serré, Amyot-Bourgeois, och Astles (2021) som analyserade agentbaserade simulerade scenarier med XAI. Mer specifikt undersöktes hur tillämpningen av SHAP (avsnitt 4.4.1) bidrar till analysen i efterarbetet med simuleringarna för ett militärt operativt scenario. Simuleringarna modellerades med ensembelmetoden *random forest*[37] efterföljt med tillämpning av SHAP och tillhörande visualiseringar. Bland slutsatserna konstateras att metodiken bistod med grundläggande förklaringar av en komplex parameterrymd som kan ligga till grund för vidare fördjupade studier.

4.6 Framtidsprognos och analys

En återkommande åsikt i litteraturen är att de tekniska framstegen behöver föregås av vad som liknas vid en filosofisk frågeställning; *vad är en förklaring?* Vad som är ställt bortom rimligt tvivel är att fältet hindras av avsaknaden av enhetliga definitioner av centrala begrepp och önskvärda egenskaper. Tvärtom skulle fältet gynnas av framtagandet av sådana definitioner. Att dessa problem kvarstår är en stark indikator på fältets övergripande omognad. Det är samtidigt noterbart att inom vissa områden har metoderna redan rönt framgångar och civila tillämpningar bedöms sannolikt överföringsbara till den militära domänen.

Genom avskanningsarbetet har det observerats en skillnad mellan begreppen förståelse och tillit. Möjligen är det en aspekt att värdera i diskussionen kring förklarbarhet baserat på användarmålgrupp. En återkommande problematisering är säkerställandet av noggrannheten för post-hoc metoder till de AI-modeller som förklaras. Detta hinder löses med transparenta, självförklarande modeller, men än så länge vanligen genom avkall på prestanda och med varierande grad av faktisk förklarbarhet. Över tid bedöms XAI som en potentiell framgångsfaktor genom att möjliggöra implementering av AI där enbart traditionella system tidigare varit anförtrodda i att komplettera människor i deras uppdrag.

[37] https://en.wikipedia.org/wiki/Random_forest (besökt mars 2023)

5 Systemperspektiv

System som innehar någon form av mjukvarukomponent i form av en modell som tränats baserat på data, dvs. där artificiell intelligens (AI) och maskininlärning (ML) använts, blir i och med de stora framsteg som gjorts inom AI på senare år alltmer intressant. En rad olika frågor kan tänka ställas vad gäller sådana system. På vilket sätt skiljer de sig från mer traditionella mjukvarusystem? Vad är den typiska karaktäristiken hos dessa system samt vilka speciella utmaningar existerar när sådana system utvecklas? Mycket forskning har framförallt skett vad gäller utveckling av AI på modellnivå såsom att identifiera och utvärdera olika modellstrukturer samt träningsalgoritmer, dvs. processen där man med hjälp av algoritmer fångar upp mönster i data på ett så bra sätt som möjligt (givet olika utvärderingsmetoder). Det främsta målet vid modellutveckling är ofta att uppnå så högt värde som möjligt för olika mått som på olika sätt mäter kvalitén av modellens prediktion för ny data, dvs. data som inte använts för träning. Vidare är många sådana modellstrukturer i dagsläget även ytterst komplexa vilket gör att ordet *arkitektur* används och där det finns forskningsområden som syftar till att söka efter lämpliga arkitekturer (Elsken, Metzen, och Hutter 2019) för problemet och data som man förfogar över.

Trots ovanstående komplexitet så är det oftast inte tillräckligt att ha en komplex modell med en given kvalité av prediktioner för att framsteg inom AI-forskningen ska komma tillämpning till gagn eftersom det också behövs komponenter och moduler runt AI-modellen som innehar ett slags utanförperspektiv på helheten där denna används, dvs. ett systemperspektiv. En AI-modell, i nuvarande terminologi, är ju endast tränad med hjälp av tidigare data som sannolikt är begränsade till en viss del eller tidsperiod av en miljö och modellen bör således inte användas i situationer som inte riktigt passar nuvarande modell eller där miljön gradvis ändrats över tid, ett fenomen som kallas konceptförskjutning (eng. concept drift) (Lu m.fl. 2018). AI-modeller syftar ofta till att på något sätt användas i en föränderlig omvärld vilket då medför att även karaktäristiken hos ny inkommande data ändras. Modellen i sig har då ingen funktion för att avgöra att den själv i någon mening är inaktuell. Det som krävs är en komponent som, möjligtvis genom att använda modellen själv, avgör att inkommande data inte längre passar nuvarande modell varvid en insats måste göras, t.ex. att modellen måste tränas om med nuvarande, eller i alla fall senare, data.

I detta kapitel redogör vi för olika aspekter som är viktiga ur en systemperspektivvinkel vid utveckling samt användning av AI-system. Inledningsvis, i avsnitt 5.1, så redogör vi för forskningsläget för AI-system från olika perspektiv. I avsnitt 5.1.1 så tas karaktäristik samt väsentliga skillnader mellan AI-system och mer traditionella mjukvarusystem upp. Sådan karaktäristik och skillnader ligger sedan till grund för senare avsnitt, dvs. avsnitt 5.1.2, som belyser utmaningar för AI-system. I avsnitt 5.1.3 redogörs för olika utvecklingsprocesser samt reglering av AI-system som syftar till att främja god kvalité och säker användning. Till sist, i avsnitt 5.2 sammanfattar vi området och beskriver framtida prognos kring de aspekter vi tagit upp i kapitlet.

5.1 Forskningsläget

I nedanstående underavsnitt redogörs för forskningsläget av AI-system utifrån tidigare nämnda aspekter. Redogörelsen är baserad på litteratursökning via Google Scholar där framförallt sökning har skett för de senaste fem åren men där också artiklar har följts genom citeringar varvid även äldre artiklar återfinns. Termer som använts för sökning är ”Artificial Intelligence Systems”, ”Machine Learning Systems” men även termer kopplade till mjukvaruutveckling inom AI/ML såsom uttrycket ”MLOps” som blivit ett etablerat uttryck för mjukvaruutveckling inom maskininlärning och artificiell intelligens samt även termer som ”Machine Learning Software” och ”Artificial Intelligence Software”.

5.1.1 Karaktäristik

En av huvudskillnaderna mellan ett mer traditionellt mjukvarusystem gentemot ett AI-system är att den senare i någon mening är mer implicit uttryckt samt mer icke-deterministisk (Giray 2021) eftersom funktioner inte är direkt utryckt (Lwakatare m.fl. 2020) i form av programkod (Martínez-Fernández m.fl. 2022). Traditionella mjukvarusystem är direkt programmerade för funktionalitet som tar hand om olika situationer eller systemtillstånd som uppstår. AI-system däremot innehar en komponent som tränats med hjälp av en viss typ av algoritm och data, dvs. själva AI-modellen, där systemfunktionalitet sedan uttrycks med hjälp av denna modell. En sådan komponent innehåller ett stokastiskt inslag som gör att ett annat förhållningsätt behövs vid både utveckling samt användning av systemet. Ett träningsförfarande med hjälp av en modell har oftast använts just för att det är svårt att direkt utrycka via programkod det som modellen predikterar (Sculley m.fl. 2015). Inom bildanalys t.ex., givet att bra träningsdata finns, kan en modell tränas för hur ett visst objekt kan se ut medan att direkt uttrycka utseendet av ett sådant objekt genom programkod utan en modell kan vara problematiskt. Den stokastiska komponenten i AI-system består främst i att den tränade modellen inte alltid kan göra felfria prediktioner (Mikkonen m.fl. 2021) och/eller resultatet ut från modellen består av en prediktion i form av en osäkerhetsrepresentation såsom en sannolikhetsfunktion. Prediktionsfel, eller osäkerhet, måste då antas finnas i systemet och kringliggande mjukvarukomponenter i systemet behöver anpassas till ett sådant förhållande. Träningsförfarandet av en modell, istället för att direkt skriva kod för motsvarande funktionalitet, har också en utgångspunkt i att komplexiteten i problemen där ett sådant förfarande väljs är hög. T.ex. för att identifiera en viss typ av objekt i en bild, och trots att man då använder en modell för sådana problem, krävs det ofta mycket kunskap och experimenterade i form av prototyper för att uppnå en tränad modell av tillräcklig kvalité för att kunna användas i en verklig tillämpning. Även om själva modellkoden i förhållande till hela kodmängden i ett system ofta är mycket liten (Sculley m.fl. 2015) så är denna del av mjukvaruprocessen krävande och annorlunda än mer traditionell mjukvaruutveckling (eng. software engineering) vilket är något som uppmärksammats i ett flertal nyligen publicerade översiktsartiklar (Giray 2021; Amershi m.fl. 2019; Martínez-Fernández m.fl. 2022; Paleyes, Urma, och Lawrence 2022). Avgörande för hur väl man lyckas fånga ett mönster via en modell beror både på modellstruktur och på träningsalgoritm samt hur väl de data man erhållit representerar den miljö i vilken systemet syftar till att användas.

Vikten av bland annat modell och träningsalgoritm speglas också i en relativt nyligen framtagen taxonomi kring olika fel som kan uppstå i djupinlärningssystem (Humbatova m.fl. 2020). Eftersom en av modellens främsta tillgångar är just data så behöver man noggrant hantera och designa system utifrån vilken typ av data som kan tänkas behövas. Men som har belysts (Sambasivan m.fl. 2021) uppfattas inte databearbetning som ett lika intressant område som modelleringsområdet vilket kan få till följd att det inte läggs tillräckligt med resurser på det förstnämnda området. Detta kan då leda till sämre övergripande kvalité av ett AI-system.

5.1.2 Utmaningar

På grund av att AI-modeller har ovanstående nämnda egenskaper så medför detta ett antal utmaningar vad gäller utveckling och användandet av sådan mjukvara i olika former av system. Det finns uppskattningar[38] kring att endast en knapp majoritet av alla AI-prototyper i slutändan når produktionsstatus vilket vidare kan ses som en indikation på svårigheter kring utveckling av AI-system som i realiteten används.

[38] https://www.gartner.com/en/newsroom/press-releases/2020-10-19-gartner-identifies-the-top-strategic-technology-trends-for-2021 (besökt maj 2023)

5.1.2.1 Underhåll

En av de uttalade svårigheterna i AI-system är att komplexiteten inte primärt ligger i dess initiala utveckling eller igångsättningsfas (eng. deployment) utan snarare i senare skede vid underhåll av sådana system (Sculley m.fl. 2015) där så kallad teknisk skuld (Cunningham 1992) (eng. technical debt) uppstår framförallt på systemnivå i gränssnitt mellan olika systemkomponenter. Något som tas upp och stöttar den tesen är specifika designmönster som ofta uppstår inom utvecklande av AI-system (Sculley m.fl. 2015). Ett exempel på ett sådant mönster är så kallad sammanfogningskod (eng. glue code) som härstammar från ett välkänt sätt att distribuera AI-modeller (metoder) via självständiga mjukvarumoduler (kan benämnas olika såsom paket eller ramverk) med givna gränssnitt och som ofta resulterar i att man måste anpassa en hel del kod till dessa gränssnitt. Sådan sammanfogningskod kan därmed utgöra ett besvärligt beroende i ett AI-system. Vidare så tas begrepp såsom ”pipeline jungles” upp som relaterar till koncept inom mjukvaruutveckling som innebär att man ofta har långa serier med sekventiell applicering av olika funktioner, ofta som en del av databearbetning, som då kan vara svåra att överskåda. Kartläggning av olika typer av teknisk skuld samt även nya typer av sådana skulder som är mer unika för AI-system, såsom modell- och dataskuld, har också genomförts (Bogner, Verdecchia, och Gerostathopoulos 2021). Även empiriska studier på ramverk för området djupinlärning (LeCun, Bengio, och Hinton 2015) visar att sådana ramverk innehåller ett flertal olika typer av tekniska skulder (J. Liu m.fl. 2020) samt antaganden (Yang m.fl. 2021) som i förlängning kan leda till teknisk skuld. Data som ligger till grund för dessa studier kommer från olika delar av ramverkens utvecklingsmiljö, t.ex. kommentarer i källkoden. Slutsatser som dras (J. Liu m.fl. 2020) är att den snabba utvecklingen som varit inom AI-området kan medföra att utvecklare tar medvetna genvägar i utvecklingen vilket då blir en teknisk skuld som på sikt kan behöva betalas av i form av framtida programmeringsresurser. Vidare så identifieras också att teknisk skuld är högre jämfört med annan typ av mjukvaruprojekt.

Problem i underhållsfasen av AI-system har även belysts från etablerade begrepp inom mjukvaruutvecklingsområden såsom modularitet, testbarhet, återanvändningsbarhet, analyserbarhet samt förändringsbarhet (Mikkonen m.fl. 2021). En till stor grad gemensam faktor som återfinns, implicit och/eller explicit, gentemot ett flertal av dessa aspekter är den stokastiska komponenten som härstammar från både användning av en AI-modell eller som en konsekvens av träning. Många träningsalgoritmer innehåller en stokastisk komponent som en del av själva träningen (M. D. Hoffman och Gelman 2014; LeCun, Bengio, och Hinton 2015) vilket gör att resultatet av träningen kan bli olika för olika träningsomgångar. Även om man kan konfigurera (pseudo-) slumptalsgeneratorer att bete sig enligt ett givet initialt mönster (så kallat ”frö”) så kan det tränade resultatet bli relativt annorlunda som en konsekvens av att man tränar på delvis ny data (Mikkonen m.fl. 2021), framförallt om AI-modellen utgörs av en modelltyp vars syfte är att fånga mönster på en generell nivå som t.ex. är fallet med djupinlärning (LeCun, Bengio, och Hinton 2015). Detta kan ske eftersom sådana modeller är uppbyggda på ett sätt där de olika inbyggda parametrarna inte korresponderar mot särskilda mönster. En annan aspekt att beakta med denna typ av modeller är svårigheten att analysera dessa (Mikkonen m.fl. 2021), vilka syftar till att fånga in mönster generellt istället för mer fördefinerade mönster via någon datagenereringsmodell eller specifika särdrag (eng. features). Komplexitet hos sådana modeller har även ökat på senare tid och moderna djupinlärningsmodeller börjar i större utsträckning även likna system[39], i alla fall ur en komplexitetsynvinkel, eftersom olika typer av block interagerar på ett komplext sätt (Vaswani m.fl. 2017).

5.1.2.2 Kontinuerligt lärande

En annan utmaning som också relaterar till underhåll är hur väl själva modellen stämmer överens med den miljö där den används. Detta benämns olika i litteraturen såsom *färskhet*

[39] På modellnivå används dock inte termen ”system” utan ”arkitektur”.

(eng. *freshness*) (Giray 2021), *kontinuerligt lärande* (eng. *continuous learning*) (Kreuzberger, Kühl, och Hirschl 2023) samt *övervakning* (eng. *monitoring*) (Symeonidis m.fl. 2022; Hendrycks m.fl. 2021) och relaterar till att den process som studeras med tillhörande data normalt kan förändras över tid och modellen behöver därför uppdateras, dvs. tränas om på ny data som är mer representativ för den aktuella miljön. En modell som är utdaterad kan medföra olika typer av konsekvenser på systemnivå. Som ett exempel kan nämnas att om en sådan modell skulle användas inom ett avvikelseupptäcktssystem (eng. anomaly detection system) så skulle man troligen få många varningar som är relativt ointressanta vilket i sin tur kan leda till att en operatör på sikt i realiteten slutar använda systemet. Detta eftersom avvikelseupptäckt kan bygga på att upptäcka data som inte längre passar modellen vilket är fallet om en miljö har ändrats och ny typ av data uppstått. Tekniskt sett så benämns problemet med inaktuell modell för konceptförskjutning och är i sig ett relativt stort forskningsområde (Bayram, Ahmed, och Kassler 2022; Hu, Kantardzic, och Sethi 2020). Problemet är också ett bra exempel på hur andra moduler och komponenter än AI-modellen är nödvändiga i ett AI-system just för att hantera perspektiv som ligger utanför modellen. Detta eftersom det krävs att man använder aktuell modell på nya inkommande datapunkter för att upptäcka att den inte längre passar den aktuella situationen.

5.1.2.3 Databearbetning

Problematiken kring konceptförskjutning kopplar också starkt an till hur man hanterar och aktivt arbetar med data. Hantering och bearbetning av data har dock belysts varit underordnat själva modelleringsprocessen (Sambasivan m.fl. 2021) vilket kan tyckas märkligt i och med den moderna AI:ns starka beroende av just data. Traditionellt så innebär hantering av data en viss bearbetning, ofta kallat databearbetning (eng. data pre-processing), och ses som ett nödvändigt steg för att ta sig vidare till modelleringssteget men som också kan vara helt avgörande för efterföljande kvalité av modelleringsarbetet. Det starka beroendet mellan dagens AI och data kan alltså ses som en begränsning. Detta har lett till områden där man försöker att balansera mellan träningsdata och befintlig kunskap som komplement (Von Rueden m.fl. 2021). En annan aspekt kring data är att den blivit en faktor för hur vissa system är uppbyggda t.ex. som distribuerade system (Q. Li m.fl. 2021) (ett område som benämns *federated learning*) eller där man strävar att bearbeta data så nära källan som möjligt (Murshed m.fl. 2021) (eng. edge computing). Vid sådana system behöver ofta flera aspekter balanseras såsom t.ex. reducering av komplexitet i modeller för att kunna användas med mindre kraftfulla datorenheter nära datakällan.

5.1.2.4 Användaraspekter

Inte bara tekniska aspekter utgör en utmaning inom AI-system, utan även en rad olika användaraspekter. Trots att det ligger inbyggt i begreppet AI att det handlar om system som till stora delar agerar autonomt så är ofta användaren av sådana system involverade på olika sätt i beslutsprocessen. I vilken omfattning sådan användarinteraktion finns kan bero på olika faktorer inom tillämpningen exempelvis konsekvenser vid felaktigt beslut. Inom medicinsk AI t.ex. (Jin m.fl. 2020) är risknivån generellt hög. Även inom detta användaraspektområde sticker den stokastiska komponenten ut inom AI-system. Hur ska en användare ta sig an problematiken kring att AI-system potentiellt kan göra felaktiga prediktioner? Komplexiteten i AI-system gör att det är ytterst problematiskt, framförallt för en användare som inte själv är utvecklare av AI, att förstå den aggregerade osäkerheten i ett sådant system och dess potentiella konsekvenser. Så kallad förklarbar AI (Arrieta m.fl. 2020) (eng. explainable AI, XAI[40]) syftar till att förklara hur en avancerad modell har kommit fram till sin prediktion och kan därmed stötta en användare i denna process men, som belysts i tidigare text, det kan finnas andra situationer där det behövs ett annat perspektiv, såsom vid konceptförskjutning (Lu m.fl. 2018) då sådana förklaringar för modellens resonemang sannolikt inte är tillräckliga. Utöver detta så kan det finnas andra

[40] Se även kapitel 3

aspekter från användarens, och även andra aktörers, sida såsom högt ställda förväntningar på systemet vilket då kan behöva justeras (Kocielnik, Amershi, och Bennett 2019) till det faktiska systemets kvalitetsnivå. Sådana förväntningar kan uppstå inte minst av de stora framsteg som gjorts inom AI-området på senare tid som även inneburit ett stort media- och allmänhetsintresse. Risken finns då att goda resultat från ett område förs över i förväntningar på ett annat område utan hänsyn till rådande förutsättningar till det senare, t.ex. brist på data som är ju är helt avgörande för nuvarande AI-systems kvalité. Som har belysts (Xu m.fl. 2023) så kan även AI-system innebära en stor förändring av hur man faktiskt jobbar med digitala verktyg, t.ex. beslutsstöd, från att ha varit mer av typen funktioner som svarar på en given användarinstruktion till att vara mer som en samarbetspartner vilket också ställer andra krav på en användare av ett sådant system.

5.1.3 Kvalitetssäkring

För att ta sig an ovanstående specifika utmaningar som uppstår inom utveckling av AI-system så har relativt mycket forskning belyst olika processer som ämnar uppnå en viss övergripande kvalité hos det utvecklade systemet. Sådana processer kopplar även starkt an till kommande reglering där dokumentation är en viktig del av AI-system.

5.1.3.1 Utvecklingsprocesser

Processer kan vara olika detaljerade och innehålla olika många steg beroende på i vilken typ av sammanhang systemet ska användas, t.ex. om det är ett säkerhetskritiskt sammanhang. Gemensamt för många av processerna, ofta under benämningen maskininlärningsoperationer (eng. machine learning operations, MLOps) (Symeonidis m.fl. 2022) är att de belyser vikten av samarbete mellan aktörer som innehar olika kompetenser eller roller, såsom t.ex. dataanalytiker (eng. data scientist) och mjukvaruutvecklare (eng. software engineer) etc. (Kreuzberger, Kühl, och Hirschl 2023) samt att detta sker kontinuerligt i mer eller mindre sammanflätade utvecklingsiterationer (Symeonidis m.fl. 2022). En aspekt att beakta utifrån ovanstående processer är att de är resurskrävande både vad gäller att en organisation måste inneha olika kompetenser samt att nivån av komplexitet även ställer stora krav på dessa kompetenser. En annan aspekt mer utifrån sammanflätningen av olika iterationer är att det uppstår en hel del beroenden mellan olika delar, t.ex. om inte databearbetningsdelen fungerar tillfredställande så kan det vara svårt att arbeta med de andra sammanlänkade stegen och således som helhet uppnå en god implementering av en utvecklingsprocess. Problem med just databearbetning samt att det är de organisationer som överlag har större personalresurser med olika slags kompetenser som kommit längst vad gäller processer såsom MLOps är något som belysts (Mäkinen m.fl. 2021).

Det finns även ett intresse av att minimera tiden det tar att fullfölja en utvecklingsiteration och en möjlighet som belysts (Symeonidis m.fl. 2022) är att automatisera själva modellutvecklingsdelen genom så kallad automatisk maskininlärning (eng. AutoML) (He, Zhao, och Chu 2021) vilket är en metodik som bygger på att algoritmer tränar en uppsättning av olika modeller varvid dessa utvärderas utifrån något eller några givna mätbara kvalitéer och där den bästa av dessa modeller sedan används som resultat. Detta är dock en resurskrävande process i termer av beräkningskraft och valet av en sådan strategi beror också på typ av modell som används. T.ex. om modellstrukturen är en generell struktur för att fånga upp mönster såsom med modeller av typen djupinlärning (eng. deep learning) (LeCun, Bengio, och Hinton 2015) jämfört med om modellen är avsiktligt designad utifrån antaganden kring hur data generats, och således även mönster i data, baserat på en mer statistiskt orienterad maskininlärning (Gelman m.fl. 2020; Gabry m.fl. 2019).

5.1.3.2 Ramverk och reglering

För att uppnå och fastställa en viss önskad kvalité av ett system har det även utvecklats mer kompletta ramverk som både innehåller uppgifter som ska genomföras i en process men även utvärdering med avseende på teknisk mognadsgrad i termer av nivåer (Lavin m.fl. 2022). Ramverket bygger på ingenjörsprinciper och syftar till att bland annat uppnå

tillförlitlighet samt robusthet genom ett antal nivåer där varje nivå beskriver vilket arbete som utförs, inklusive en specifik beskrivning av hur man arbetar med data, samt hur man utvärderar olika aspekter på just den nivån. Inom ramverket belyses också vikten av dokumentation för de olika nivåerna, något som även har tagits upp i samband med förslag kring nytt regelverk inom EU för framförallt högrisk AI-system (Mökander m.fl. 2022). Att arbeta systematiskt med design, antaganden, och utvärdering av AI-system samt dokumentera alla sådana delar av processer kommer där sannolikt vara en förutsättning för att kunna utvärdera huruvida ett system uppfyller ett visst regelverk inom AI. Även efter att ett visst system tagits i bruk inom ett visst område så kommer det även sannolikt finnas krav på uppföljning under efterföljande livscykel.

5.2 Framtidsprognos och analys

AI-modeller är sannolikt något som kommer vara alltmer förekommande i de flesta tillämpningar och system då det kan innebära en förenkling samt ge ökad produktivitet inom en rad olika områden. Mjukvara har tagit en allt större plats inom många sektorer av samhället och spelar många gånger en avgörande roll för hur en produkt uppfattas. Utifrån denna synvinkel kan man se AI-system som ett nästa led inom mjukvarusystem där det då finns potential för en större grad av autonomi och hantering av mer komplexa situationer eller information jämfört med tidigare system. Som har belysts i avsnitt 5.1 så finns det dock stora skillnader mellan AI-system och mer traditionella mjukvarusystem, både vad gäller utveckling och användande. För ökad förståelse behöver erfarenheter av utveckling och användning av AI-system ackumuleras över tid. Utvecklingshastigheten för nuvarande AI-metoder har varit hög och att inkorporera sådana metoder i praktiska sammanhang tar tid speciellt med avseende på hur den stokastiska komponenten kan hanteras. Mycket handlar också om att bygga upp erfarenhet kring att arbeta med data och utveckling av modeller för ett syfte som sträcker sig bortom prototypstadiet. Mer forskning samt utvärdering kring hur man bäst tar sig an sådana processer är även då nödvändigt eftersom de flesta utvecklingsprocesser är relativt nya och snarare ett resultat av olika erfarenheter jämfört med mer utvärderingsorienterade studier. Regelverk för olika AI-system är under utveckling och det kommer sannolikt ta lång tid att inhämta kunskap samt utvärdera vilka processer etc. som lämpar sig bra för en viss typ av regelefterlevnad eller önskvärd kvalitet av ett visst system. Med tanke på att AI och ML länge präglats av modeller och algoritmer och inte i lika stor utsträckning systemperspektiv så kommer sannolikt ett sådant perspektiv få än mer uppmärksamhet i framtiden, inte minst på grund av att det är så AI-modeller kommer olika tillämpningar till gagn. Således är området AI-system ett högst relevant, nödvändigt område, men där mer forskning och utvärdering behövs för att säkerställa processer och rutiner för utveckling och underhåll av sådana system för att det ska uppfylla önskvärda egenskaper, t.ex. tillförlitlighet, över tid.

6 Värdering och slutsatser

Rapporten omfattar en avskanning under början av 2023 av några delområden inom dataanalys och AI närmare bestämt preskriptiv analys (kapitel 2), osäkerhetshantering (kapitel 3), förklarbarhet (kapitel 4) och systemperspektiv (kapitel 5). Nedan följer en kort sammanfattning av de viktigaste slutsatserna för de studerade områdenas och jämförande visualisering.

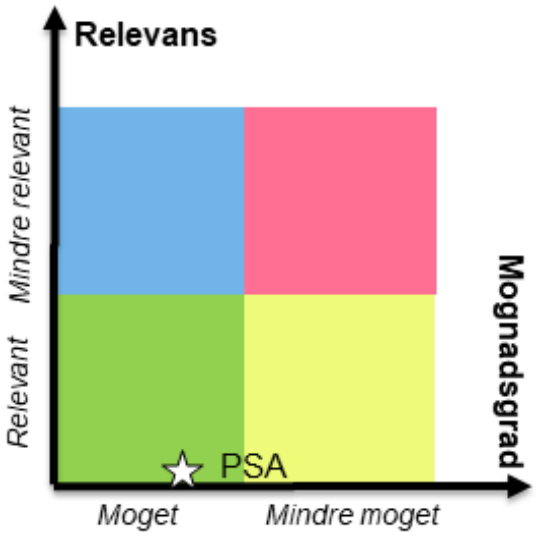
Vi värderar delområdena och vissa metoder med avseende på två kriterier: *mognad* och *relevans*. Med mognad avses hur välutvecklat materialet i en artikel är i termer av möjligheten att bidra till praktiska tillämpningar (dvs. i vilken grad tekniken är utvecklad). Försvårande i sammanhanget kan vara att en viss algoritm mest kan hantera förenklade problem, eller att tillämpningen kräver svåråtkomlig data eller expertkunskaper. Relevans handlar om i vilken utsträckning materialet i forskningsartikeln är nyskapande, dess potential och i vilken utsträckning den kan tänkas bidra till försvarstillämpningar.

De avskannade områdenas och metoderna placering i diagrammen indikeras med en ☆-symbol. Ju närmare origo en artikel placerats desto mer intressant är det ur ett försvarstillämpningsperspektiv.

6.1 Preskriptiv analys

Preskriptiv analys (PDA) är en form av dataanalys som fokuserar på beslutsfattande baserat på data och analys av data. Att fatta effektiva beslut grundade på data, är inte bara relevant i militära sammanhang, det är nödvändigt och livsviktigt i en militär operation. Det skall också användas till myndighetens övriga verksamhet (precis som för företag och organisationer i övrigt), där det är viktigt för att säkerställa att vi får ut maximal effekt av de resurser Försvarsmakten (FM) förfogar över.

PDA för att optimera verksamheten i (civila) organisationer har potential att vidareutvecklas, vilket även FM kan dra nytta av. Operationsanalys och optimering är en del av PDA, och väl använda metoder inom militären. På så sätt är det inget nytt. Det som är nytt är optimalt användande av de stora datamängder som idag och i framtiden är tillgängliga för militärt beslutsfattande. I detta avseende finns stora möjligheter till ökad användning av PDA. Analysen sammanfattas i figur 9.

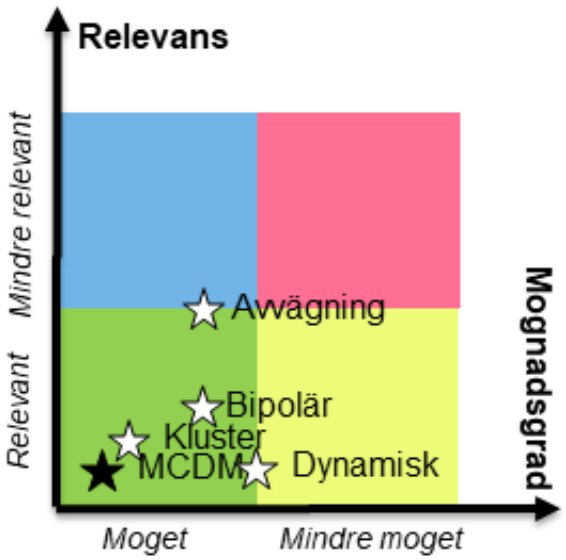


Figur 9. Området preskriptiv analys (PDA) värderas för försvarstillämpningar

Ett område som bidrar till PDA är multikriterieanalys (MCDM, avsnitt 2.1). En militär chef skall fatta beslut i situationer där många faktorer påverkar beslutet, men de kan vara svåra eller omöjliga att väga eller jämföra, varvid MCDM är användbart.

Att använda en strukturerad och systematisk metod som MCDM för militärt beslutsfattande ökar förståelsen för beslutets grunder, och minskar risken för mänskliga bias, och kan således ta fram ett bättre beslut jämfört med att endast använda en mänsklig bedömning. Metoden är användbar och relevant både inom t.ex. resursallokering och insatsplanering såväl som inom upphandling och försvarsplanering.

MCDM är ett område som har rötterna i mitten av förra århundradet och som i sig bedöms generellt som relativt moget. Det finns dock fortfarande utvecklingsspår utöver de etablerade inom MCDM som är mindre mogna, t.ex. beträffande dynamiskt beslutsfattande. Generellt är det en kort väg från framtagande av nya teorier och algoritmer till tillämpning på praktiska problem. Analysen sammanfattas i figur 10.

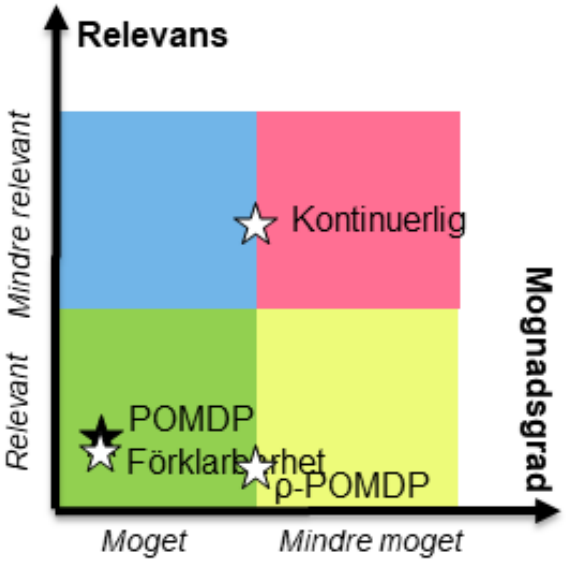


Figur 10. I figuren jämförs metoder inom MCDM med avseende på relevans och mognad

POMPD (avsnitt 2.2) är en annan metod som möjliggör beslutsfattande. Militärt beslutsfattande kännetecknas av osäkerhet i utfall och avsaknad av fullständig information kring fienden (dvs ”krigets dimma”), vilket POMDP hanterar.

POMDP är därmed användbara och relevanta i en mängd olika militärt relevanta situationer, inklusive styrning av robotar och autonoma farkoster, styrning och resursallokering av sensorer för optimal informationsinhämtning, cyberförsvar, framtagande av planer, för resursfördelning och val av mål.

POMDP som teknik är generellt mogen och välanvänd i en mängd problem, t.ex. robotik och resursallokering. Det är dock inte undersökt hur mycket POMDP faktiskt används för framtagande av militära planer och liknande. Generellt är det ganska kort väg från framtagande av nya teorier och algoritmer till tillämpning på praktiska problem. Analysen sammanfattas i figur 11

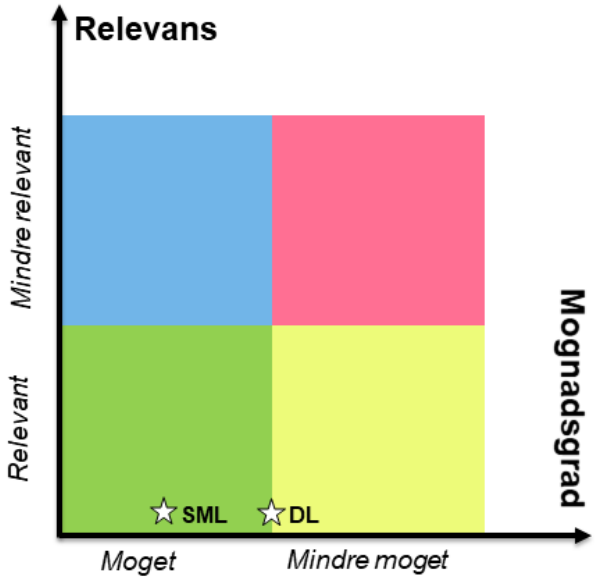


Figur 11. I figuren jämförs metoder inom POMDP med avseende på relevans och mognad

6.2 Osäkerhetshantering

I rapporten jämförs osäkerhetshantering inom två typer av (databaserad) maskininlärning, djupinlärning (DL) och statistisk maskininlärning (SML). DL-metoder (baserat på djupa neuronnät) skiljer sig markant från statistisk maskininlärning genom att de är konstruerade för att fånga upp mönster på en generell nivå istället för som i fallet för statistisk maskininlärning där datagenereringsprocessen ofta antas följa sannolikhetsfördelningar och således görs antaganden om mönsterformerna. Denna skillnad, tillsammans med modellstorlek och datamängd, är avgörande för de träningsalgoritmer som oftast används inom de olika områdena. Inom DL blir det då oftast någon form av optimeringsorienterad träningsalgoritm som används där kvalitén av osäkerhetsmodelleringen (om en sådan modellering existerar) kan vara svår att bedöma medan man inom SML använder mer av samplingsbaserade metoder som ofta har olika typer av kvalitetsgarantier kring osäkerhet givet att vissa villkor är uppfyllda. Som framgår i de båda avsnitten i rapporten gör detta att DL-området blir mer spretigt vad gäller metoder eftersom det där ofta handlar om olika typer av heuristiska ansatser för osäkerhetsmodellering jämfört med SML där ofta samplingsansatsen är en naturlig del. Dock så gäller det senare i fallen där datamängderna är rimliga då samplingsansatsen kan vara beräkningsmässigt utmanande.

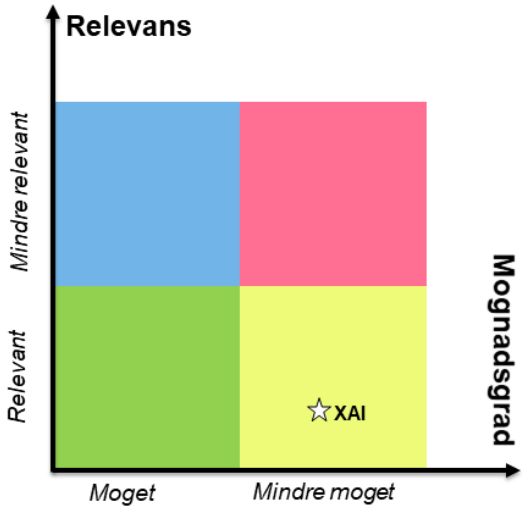
Då SML är mer av en generell grund för osäkerhetshantering är det ett mer moget och relevant område för osäkerhetsmodellering inom dataanalys, dock med vissa begränsningar vad gäller storskaliga datamängder och komplexa modeller medan med DL kan området gagnas av ytterligare studier för att uppnå mer kunskap kring vilken träningsansats och modell som lämpar sig för bra modellering av osäkerhet. Båda områdena är högst relevanta, speciellt inom områden där det finns en riskkomponent såsom i säkerhetskritiska system, dvs. där fel beslut kan leda till stora negativa konsekvenser. Analysen sammanfattas i figur 12.



Figur 12. I figuren jämförs osäkerhetshanteringsmetoder inom statistisk maskininlärning (SML) och djupinlärning (DL) med avseende på relevans och mognad

6.3 Förklarbarhet

Området Förklarbarhet (XAI) är generellt sett omoget, till viss del på grund av att begreppet inte ens är tillräckligt väl förstått. Förklarbarhet är ett generellt begrepp och rör förklaring av alla möjliga typer av AI-metoder, men det aktuella behovet är starkt kopplat till djupa neuronnät vars beteende normalt sett är svårt att förklara. Samtidigt är området av intresse för försvarstillämpningar vilket inte minst Darpas engagemang signalerar. Analysen sammanfattas i figur 13.



Figur 13. Området förklarbarhet (XAI) värderas för försvarstillämpningar med avseende på relevans och mognad

6.4 Systemperspektiv

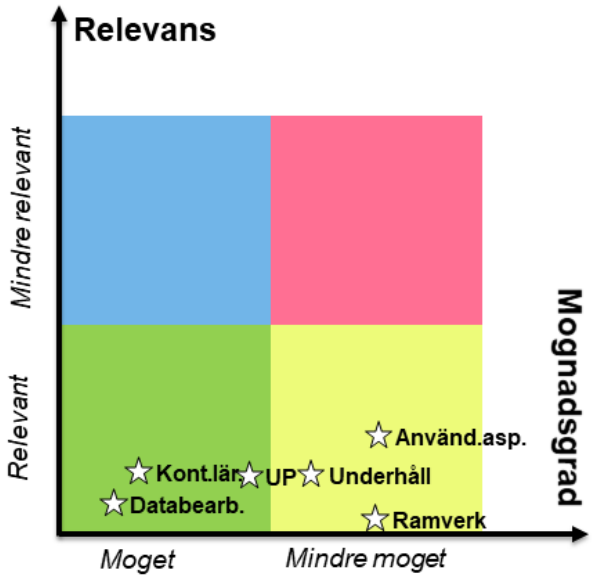
I kapitel 5 presenteras ett antal systemperspektiv på AI-system. Den följande analysen sammanfattas i figur 14. Kontinuerligt lärande (betecknat ”Kont.lär.” i figuren) och databearbetning (”Databearb.”) är etablerade områden. För alla moderna AI-system är data och databearbetning centralt därav hög relevans, men som tas upp i avsnitt 5.1.2 kan mer fokus

behöva läggas på databearbetning så att det inte bara ses som ett delsteg på vägen till inlärning. Beroende på AI-system, t.ex. system som verkar inom en dynamisk föränderlig miljö, så innehar kontinuerligt lärande olika relevans, men det är en aspekt från systemperspektiv som alltid bör beaktas och därav hög relevans.

Underhåll (”Underhåll”) och utvecklingsprocesser (”UP”) är relativt nya områden med avseende på AI-system. Områdena hänger ihop, en väl dokumenterad process ger upphov till en medvetenhet kring t.ex. teknisk skuld och då vilka delar av systemet behöva fokuseras på vid underhåll och utveckling. Vad gäller utvecklingsprocesser och mognadsgrad så är sannolikt mer av framtida erfarenheter och utvärdering av olika processer något som behövs och detta gäller till stor del även underhåll. Båda områdena är i någon mening lika relevanta.

Användaraspekter (”Använd.asp.”) om hur människor bäst använder AI-system, eller kanske snarare samarbetar med dem, behövs det sannolikt ackumuleras mer erfarenhet kring samt ytterligare studier bedrivas. Om systemet har en hög grad av mänsklig inblandning, dvs. AI-systemet är mer av beslutsstödsystem, så har området hög relevans.

Ramverk och regleringsförslag (”Ramverk”) för AI-system finns men hur sådana förslag i praktiken bäst realiseras med hjälp av utvecklingsprocesser och andra ramverk återstår att se. Ramverk och regleringsförslag är relevant, särskilt för säkerhetskritiska system.



Figur 14. I figuren jämförs olika systemperspektiv på AI med avseende på relevans och mognad

7 Referenser

Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, m.fl. 2016. ”TensorFlow: a system for large-scale machine learning”. I *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 265–83.

Abdar, Moloud, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, m.fl. 2021. ”A review of uncertainty quantification in deep learning: Techniques, applications and challenges”. *Information Fusion* 76: 243–97.

Adadi, Amina, och Mohammed Berrada. 2018. ”Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)”. *IEEE Access* 6: 52138–60. https://doi.org/10.1109/ACCESS.2018.2870052.

Amershi, Saleema, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, och Thomas Zimmermann. 2019. ”Software engineering for machine learning: A case study”. I *2019 IEEE/ACM 41st international conference on software engineering: Software engineering in practice (ICSE-SEIP)*, 291–300. IEEE.

Andrychowicz, Marcin, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, och Wojciech Zaremba. 2017. ”Hindsight Experience Replay”. I *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/453fadbd8a1a3af50a9df4df899537b5-Abstract.html.

Antoniak, Charles E. 1974. ”Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems”. *The annals of statistics*, 1152–74.

Araya, Mauricio, Olivier Buffet, Vincent Thomas, och Françcois Charpillet. 2010. ”A POMDP extension with belief-dependent rewards”. I *Advances in neural information processing systems*, redigerad av J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, och A. Culotta. Vol. 23. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2010/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf.

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, m.fl. 2020. ”Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”. *Information fusion* 58: 82–115.

”Artificial Intelligence – Air Force Research Laboratory”. 2023. 2023. https://afresearchlab.com/technology/artificial-intelligence/.

Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, och Wojciech Samek. 2015. ”On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation”. *PLOS ONE* 10 (7): e0130140. https://doi.org/10.1371/journal.pone.0130140.

Bardenet, Rémi, Arnaud Doucet, och Christopher C Holmes. 2017. ”On Markov chain Monte Carlo methods for tall data”. *Journal of Machine Learning Research* 18 (47).

Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, m.fl. 2020. ”Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI”. *Information Fusion* 58 (juni): 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Bayram, Firas, Bestoun S Ahmed, och Andreas Kassler. 2022. ”From concept drift to model degradation: An overview on performance-aware drift detectors”. *Knowledge-Based Systems*, 108632.

Betancourt, Michael. 2015. ”The fundamental incompatibility of scalable Hamiltonian Monte Carlo and naive data subsampling”. I *International conference on machine learning*, 533–40.

Betancourt, Michael. 2017. ”A conceptual introduction to Hamiltonian Monte Carlo”. *arXiv preprint arXiv:1701.02434*.

Bingham, Eli, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, och Noah D Goodman. 2019. ”Pyro: Deep universal probabilistic programming”. *The Journal of Machine Learning Research* 20 (1): 973–78.

Blei, David M, Alp Kucukelbir, och Jon D McAuliffe. 2017. ”Variational inference: A review for statisticians”. *Journal of the American statistical Association* 112 (518): 859–77.

Bogner, Justus, Roberto Verdecchia, och Ilias Gerostathopoulos. 2021. ”Characterizing technical debt and antipatterns in AI-based systems: A systematic mapping study”. I *2021 IEEE/ACM international conference on technical debt (TechDebt)*, 64–73. IEEE.

Bryant, Michael, och Erik Sudderth. 2012. ”Truly nonparametric online variational inference for hierarchical Dirichlet processes”. *Advances in Neural Information Processing Systems* 25.

Cambria, Erik, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, och Navid Nobani. 2023. ”A Survey on XAI and Natural Language Explanations”. *Information Processing & Management* 60 (1): 103111. https://doi.org/10.1016/j.ipm.2022.103111.

Campbell, Trevor, och Boyan Beronov. 2019. ”Sparse variational inference: Bayesian coresets from scratch”. *Advances in Neural Information Processing Systems* 32.

Campbell, Trevor, och Tamara Broderick. 2019. ”Automated scalable Bayesian inference via Hilbert coresets”. *The Journal of Machine Learning Research* 20 (1): 551–88.

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, och Allen Riddell. 2017. ”Stan: A probabilistic programming language”. *Journal of statistical software* 76 (1).

Carrell, Annabelle, Neil Mallinar, James Lucas, och Preetum Nakkiran. 2022. ”The calibration generalization gap”. *arXiv preprint arXiv:2210.01964*.

Carvalho, Diogo V., Eduardo M. Pereira, och Jaime S. Cardoso. 2019. ”Machine Learning Interpretability: A Survey on Methods and Metrics”. *Electronics* 8 (8): 832. https://doi.org/10.3390/electronics8080832.

Chadès, Iadine, Luz V. Pascal, Sam Nicol, Cameron S. Fletcher, och Jonathan Ferrer-Mestres. 2021. ”A Primer on Partially Observable Markov Decision Processes (POMDPs)”. *Methods in Ecology and Evolution* 12 (11): 2058–72. https://doi.org/10.1111/2041-210X.13692.

Chen, Naitong, Zuheng Xu, och Trevor Campbell. 2022. ”Bayesian inference via sparse Hamiltonian flows”. *arXiv preprint arXiv:2203.05723*.

Chen, Tianqi, Emily Fox, och Carlos Guestrin. 2014. ”Stochastic gradient Hamiltonian Monte Carlo”. I *International conference on machine learning*, 1683–91.

Confalonieri, Roberto, Ludovik Coba, Benedikt Wagner, och Tarek R. Besold. 2021. ”A Historical Perspective of Explainable Artificial Intelligence”. *WIREs Data Mining and Knowledge Discovery* 11 (1): e1391. https://doi.org/10.1002/widm.1391.

Cunningham, Ward. 1992. ”The WyCash portfolio management system”. I *Addendum to the proceedings on object-oriented programming systems, languages, and applications (addendum)*, 29–30. OOPSLA ’92. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/157709.157715.

Damianou, Andreas, och Neil D Lawrence. 2013. ”Deep Gaussian processes”. I *Artificial intelligence and statistics*, 207–15.

Danilevsky, Marina, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, och Prithviraj Sen. 2020. ”A Survey of the State of Explainable AI for Natural Language Processing”. I *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 447–59. Suzhou, China: Association for Computational Linguistics. https://aclanthology.org/2020.aacl-main.46.

desJardins, Marie E., Edmund H. Durfee, Jr Charles L. Ortiz, och Michael J. Wolverton. 1999. ”A Survey of Research in Distributed, Continual Planning”. *AI Magazine* 20 (4): 13–13. https://doi.org/10.1609/aimag.v20i4.1475.

Dong, Xibin, Zhiwen Yu, Wenming Cao, Yifan Shi, och Qianli Ma. 2020. ”A survey on ensemble learning”. *Frontiers of Computer Science* 14 (2): 241–58.

Dwivedi, Rudresh, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, m.fl. 2023. ”Explainable AI (XAI): Core Ideas, Techniques, and Solutions”. *ACM Computing Surveys* 55 (9): 1–33. https://doi.org/10.1145/3561048.

Elsken, Thomas, Jan Hendrik Metzen, och Frank Hutter. 2019. ”Neural architecture search: A survey”. *The Journal of Machine Learning Research* 20 (1): 1997–2017.

Europeiska kommissionen. 2023. ”En europeisk strategi för artificiell intelligens”. 26 januari 2023. https://digital-strategy.ec.europa.eu/sv/policies/european-approach-artificial-intelligence.

Fehr, Mathieu, Olivier Buffet, Vincent Thomas, och Jilles Dibangoye. 2018. ”rho-POMDPs have lipschitz-continuous epsilon-Optimal value functions”. I *Advances in neural information processing systems*, redigerad av S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, och R. Garnett. Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/de7f47e09c8e05e6021ababdf6bc58e7-Paper.pdf.

Ferguson, Thomas S. 1973. ”A Bayesian analysis of some nonparametric problems”. *The annals of statistics*, 209–30.

Ferrer-Mestres, Jonathan, Thomas G. Dietterich, Olivier Buffet, och Iadine Chades. 2021. ”K-N-momdps: Towards interpretable solutions for adaptive management”. *Proceedings of the AAAI Conference on Artificial Intelligence* 35 (17): 14775–84. https://doi.org/10.1609/aaai.v35i17.17735.

Ferrer-Mestres, Jonathan, Thomas G. Dietterich, Olivier Buffet, och Iadine Chadès. 2020. ”Solving K-MDPs”. *Proceedings of the International Conference on Automated Planning and Scheduling* 30 (1): 110–18. https://doi.org/10.1609/icaps.v30i1.6651.

Ferson, Scott, Cliff A Joslyn, Jon C Helton, William L Oberkampf, och Kari Sentz. 2004. ”Summary from the epistemic uncertainty workshop: consensus amid diversity”. *Reliability Engineering & System Safety* 85 (1–3): 355–69.

Fortuin, Vincent. 2022. ”Priors in Bayesian deep learning: A review”. *International Statistical Review*.

Frazzetto, Davide, Thomas Dyhre Nielsen, Torben Bach Pedersen, och Laurynas Šikšnys. 2019. ”Prescriptive Analytics: A Survey of Emerging Trends and Technologies”. *The VLDB Journal* 28 (4): 575–95. https://doi.org/10.1007/s00778-019-00539-y.

Främling, Kary. 2020. ”Decision Theory Meets Explainable AI”. I *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, redigerad av Davide Calvaresi, Amro Najjar, Michael Winikoff, och Kary Främling, 57–74. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-51924-7_4.

Främling, Kary, Marcus Westberg, Martin Jullum, Manik Madhikermi, och Avleen Malhi. 2021. ”Comparison of Contextual Importance and Utility with LIME and Shapley Values”. I *Explainable and Transparent AI and Multi-Agent Systems*, redigerad av Davide Calvaresi, Amro Najjar, Michael Winikoff, och Kary Främling, 12688:39–54. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-82017-6_3.

Gabry, Jonah, och Tristan Mahr. 2022. ”bayesplot: Plotting for Bayesian models”. https://mc-stan.org/bayesplot/.

Gabry, Jonah, Daniel Simpson, Aki Vehtari, Michael Betancourt, och Andrew Gelman. 2017. ”Visualization in Bayesian workflow”. *arXiv preprint arXiv:1709.01449*.

Gabry, Jonah, Daniel Simpson, Aki Vehtari, Michael Betancourt, och Andrew Gelman. 2019. ”Visualization in bayesian workflow”. *Journal of the Royal Statistical Society Series A: Statistics in Society* 182 (2): 389–402.

Gal, Yarin, och Zoubin Ghahramani. 2016. ”Dropout as a Bayesian approximation: Representing model uncertainty in deep learning”. I *international conference on machine learning*, 1050–59.

Ganaie, M.A., Minghui Hu, A.K. Malik, M. Tanveer, och P.N. Suganthan. 2022. ”Ensemble deep learning: A review”. *Engineering Applications of Artificial Intelligence* 115: 105151. https://doi.org/10.1016/j.engappai.2022.105151.

Gawlikowski, Jakob, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, m.fl. 2021. ”A survey of uncertainty in deep neural networks”. *arXiv preprint arXiv:2107.03342*.

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, och Donald B Rubin. 2013. *Bayesian data analysis, third edition*. Chapman and Hall/CRC.

Gelman, Andrew, Aki Vehtari, Daniel Simpson, Charles C Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, och Martin Modrák. 2020. ”Bayesian workflow”. *arXiv preprint arXiv:2011.01808*.

Giray, Görkem. 2021. ”A software engineering perspective on engineering machine learning systems: State of the art and challenges”. *Journal of Systems and Software* 180: 111031.

Gunning, David, och David Aha. 2019. ”DARPA’s Explainable Artificial Intelligence (XAI) Program”. *AI Magazine* 40 (2): 44–58. https://doi.org/10.1609/aimag.v40i2.2850.

Gunning, David, Eric Vorm, Jennifer Yunyan Wang, och Matt Turek. 2021. ”DARPA’s explainable AI (XAI) program: A retrospective”. *Applied AI Letters* 2 (4): e61. https://doi.org/10.1002/ail2.61.

Guo, Chuan, Geoff Pleiss, Yu Sun, och Kilian Q Weinberger. 2017. ”On calibration of modern neural networks”. I *International conference on machine learning*, 1321–30.

Gustafsson, Fredrik K, Martin Danelljan, och Thomas B Schon. 2020. ”Evaluating scalable Bayesian deep learning methods for robust computer vision”. I *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 318–19.

Gürbüzbalaban, Mert, Xuefeng Gao, Yuanhan Hu, och Lingjiong Zhu. 2021. ”Decentralized stochastic gradient Langevin dynamics and Hamiltonian Monte Carlo”. *Journal of machine learning research* 22.

He, Xin, Kaiyong Zhao, och Xiaowen Chu. 2021. ”AutoML: A survey of the state-of-the-art”. *Knowledge-Based Systems* 212: 106622.

Hendrycks, Dan, Nicholas Carlini, John Schulman, och Jacob Steinhardt. 2021. ”Unsolved problems in ml safety”. *arXiv preprint arXiv:2109.13916*.

Heuillet, Alexandre, Fabien Couthouis, och Natalia Díaz-Rodríguez. 2021. ”Explainability in Deep Reinforcement Learning”. *Knowledge-Based Systems* 214 (februari): 106685. https://doi.org/10.1016/j.knosys.2020.106685.

Hoffman, Matthew D, David M Blei, Chong Wang, och John Paisley. 2013. ”Stochastic variational inference”. *Journal of Machine Learning Research*.

Hoffman, Matthew D, och Andrew Gelman. 2014. ”The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.” *Journal of Machine Learning Research* 15 (1): 1593–1623.

Hoffman, Robert R., Shane T. Mueller, Gary Klein, och Jordan Litman. 2019. ”Metrics for Explainable AI: Challenges and Prospects”. arXiv. https://doi.org/10.48550/arXiv.1812.04608.

Holland Michel, Arthur. 2020. ”The Black Box, Unlocked: Predictability and Understandability in Military AI”. United Nations Institute for Disarmament Research. https://doi.org/10.37559/SecTec/20/AI1.

Holzinger, Andreas, Bernd Malle, Anna Saranti, och Bastian Pfeifer. 2021. ”Towards Multi-Modal Causability with Graph Neural Networks Enabling Information Fusion for Explainable AI”. *Information Fusion* 71 (juli): 28–37. https://doi.org/10.1016/j.inffus.2021.01.008.

Hu, Hanqing, Mehmed Kantardzic, och Tegjyot S Sethi. 2020. ”No Free Lunch Theorem for concept drift detection in streaming data classification: A review”. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10 (2): e1327.

Huggins, Jonathan, Trevor Campbell, och Tamara Broderick. 2016. ”Coresets for scalable Bayesian logistic regression”. *Advances in Neural Information Processing Systems* 29.

Hult, Gunnar, Therese Almbladh, Marcus Dansarie, Johan Granholm, Eva Lagg, Stefan Silfverskiöld, och Daniel Thenander. 2022. *Technology Forecast 2022 – Military Utility of Future Technologies*. Försvarshögskolan (FHS). http://urn.kb.se/resolve?urn=urn:nbn:se:fhs:diva-11254.

Humbatova, Nargiz, Gunel Jahangirova, Gabriele Bavota, Vincenzo Riccio, Andrea Stocco, och Paolo Tonella. 2020. ”Taxonomy of real faults in deep learning systems”. I *Proceedings of the ACM/IEEE 42nd international conference on software engineering*, 1110–21.

Hwang, Ching-Lai, och Kwangsun Yoon. 1981. ”Multiple Attribute Decision Making: Methods and Applications A State-of-the-Art Survey”. I *Multiple Attribute Decision Making*. Lecture Notes in Economics and Mathematical Systems. Springer Berlin, Heidelberg.

Hüllermeier, Eyke, och Willem Waegeman. 2021. ”Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods”. *Machine Learning* 110 (3): 457–506.

Ijadi Maghsoodi, Abteen, Azad Kavian, Mohammad Khalilzadeh, och Willem K.M. Brauers. 2018. ”CLUS-MCDA: A Novel Framework Based on Cluster Analysis and Multiple Criteria Decision Theory in a Supplier Selection Problem”. *Computers & Industrial Engineering* 118 (april): 409–22. https://doi.org/10.1016/j.cie.2018.03.011.

Izmailov, Pavel, Sharad Vikram, Matthew D Hoffman, och Andrew Gordon Gordon Wilson. 2021. ”What are Bayesian neural network posteriors really like?” I *International conference on machine learning*, 4629–40.

Jiang, Xiaofeng, Jian Yang, Xiaobin Tan, och Hongsheng Xi. 2019. ”Observation-Based Optimization for POMDPs With Continuous State, Observation, and Action Spaces”. *IEEE Transactions on Automatic Control* 64 (5): 2045–52. https://doi.org/10.1109/TAC.2018.2861910.

Jin, Cheng, Weixiang Chen, Yukun Cao, Zhanwei Xu, Zimeng Tan, Xin Zhang, Lei Deng, m.fl. 2020. ”Development and evaluation of an Artificial Intelligence system for COVID-19 diagnosis”. *Nature communications* 11 (1): 5088.

Johndrow, James, Paulo Orenstein, och Anirban Bhattacharya. 2020. ”Scalable approximate MCMC algorithms for the horseshoe prior”. *Journal of Machine Learning Research* 21 (73).

Joshi, Gargi, Rahee Walambe, och Ketan Kotecha. 2021. ”A Review on Explainability in Multimodal Deep Neural Nets”. *IEEE Access* 9: 59800–821. https://doi.org/10.1109/ACCESS.2021.3070212.

Kocielnik, Rafal, Saleema Amershi, och Paul N Bennett. 2019. ”Will you accept an imperfect AI? Exploring designs for adjusting end-user expectations of AI systems”. I *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–14.

Kreuzberger, Dominik, Niklas Kühl, och Sebastian Hirschl. 2023. ”Machine learning operations (MLOps): Overview, definition, and architecture”. *IEEE access : practical innovations, open solutions*.

Kucukelbir, Alp, Dustin Tran, Rajesh Ranganath, Andrew Gelman, och David M Blei. 2017. ”Automatic differentiation variational inference”. *Journal of machine learning research*.

Lakshminarayanan, Balaji, Alexander Pritzel, och Charles Blundell. 2017. ”Simple and scalable predictive uncertainty estimation using deep ensembles”. *Advances in neural information processing systems* 30.

Lavin, Alexander, Ciarán M Gilligan-Lee, Alessya Visnjic, Siddha Ganju, Dava Newman, Sujoy Ganguly, Danny Lange, m.fl. 2022. ”Technology readiness levels for machine learning systems”. *Nature Communications* 13 (1): 6039.

LeCun, Yann, Yoshua Bengio, och Geoffrey Hinton. 2015. ”Deep learning”. *Nature* 521 (7553): 436–44.

Lee, Jaehoon, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, och Jascha Sohl-Dickstein. 2017. ”Deep neural networks as Gaussian processes”. *arXiv preprint arXiv:1711.00165*.

Lepenioti, Katerina, Alexandros Bousdekis, Dimitris Apostolou, och Gregoris Mentzas. 2020. ”Prescriptive Analytics: Literature Review and Research Challenges”. *International Journal of Information Management* 50 (februari): 57–70. https://doi.org/10.1016/j.ijinfomgt.2019.04.003.

———. 2021. ”Human-Augmented Prescriptive Analytics With Interactive Multi-Objective Reinforcement Learning”. *IEEE Access* 9: 100677–93. https://doi.org/10.1109/ACCESS.2021.3096662.

Letzgus, Simon, Patrick Wagner, Jonas Lederer, Wojciech Samek, Klaus-Robert Müller, och Grégoire Montavon. 2022. ”Toward Explainable Artificial Intelligence for Regression Models: A methodological perspective”. *IEEE Signal Processing Magazine* 39 (4): 40–58. https://doi.org/10.1109/MSP.2022.3153277.

Li, Naiqi, Wenjie Li, Yong Jiang, och Shu-Tao Xia. 2022. ”Deep Dirichlet process mixture models”. I *Proceedings of the thirty-eighth conference on uncertainty in artificial intelligence*, redigerad av James Cussens och Kun Zhang, 180:1138–47. Proceedings of machine learning research. PMLR. https://proceedings.mlr.press/v180/li22c.html.

Li, Qinbin, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, och Bingsheng He. 2021. ”A survey on federated learning systems: vision, hype and reality for data privacy and protection”. *IEEE Transactions on Knowledge and Data Engineering*.

Linardatos, Pantelis, Vasilis Papastefanopoulos, och Sotiris Kotsiantis. 2020. ”Explainable AI: A Review of Machine Learning Interpretability Methods”. *Entropy* 23 (1): 18. https://doi.org/10.3390/e23010018.

Liu, Haitao, Yew-Soon Ong, Xiaobo Shen, och Jianfei Cai. 2020. ”When Gaussian process meets big data: A review of scalable GPs”. *IEEE transactions on neural networks and learning systems* 31 (11): 4405–23.

Liu, Jiakun, Qiao Huang, Xin Xia, Emad Shihab, David Lo, och Shanping Li. 2020. ”Is using deep learning frameworks free? Characterizing technical debt in deep learning frameworks”. I *Proceedings of the ACM/IEEE 42nd international conference on software engineering: Software engineering in society*, 1–10.

Lu, Jie, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, och Guangquan Zhang. 2018. ”Learning under concept drift: A review”. *IEEE transactions on knowledge and data engineering* 31 (12): 2346–63.

Lundberg, Scott M, och Su-In Lee. 2017. ”A Unified Approach to Interpreting Model Predictions”. I *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Luotsinen, Linus, Daniel Oskarsson, Peter Svenmarck, och Ulrika Wickenberg Bolin. 2019. ”Explainable Artificial Intelligence: Exploring XAI Techniques in Military Deep Learning Applications”. https://www.foi.se/rapporter/rapportsammanfattning.html'?reportNumber='FOI-R--4849--SE.

Lwakatare, Lucy Ellen, Aiswarya Raj, Ivica Crnkovic, Jan Bosch, och Helena Holmström Olsson. 2020. ”Large-scale machine learning systems in real-world industrial settings: A review of challenges and solutions”. *Information and software technology* 127: 106368.

Maddox, Wesley J, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, och Andrew Gordon Wilson. 2019. ”A simple baseline for Bayesian uncertainty in deep learning”. *Advances in Neural Information Processing Systems* 32.

Manchingal, Shireen Kudukkil, och Fabio Cuzzolin. 2022. ”Epistemic deep learning”. *arXiv preprint arXiv:2206.07609*.

Martínez-Fernández, Silverio, Justus Bogner, Xavier Franch, Marc Oriol, Julien Siebert, Adam Trendowicz, Anna Maria Vollmer, och Stefan Wagner. 2022. ”Software engineering for AI-based systems: a survey”. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 31 (2): 1–59.

Mikkonen, Tommi, Jukka K Nurminen, Mikko Raatikainen, Ilenia Fronza, Niko Mäkitalo, och Tomi Männistö. 2021. ”Is machine learning software just software: A maintainability view”. I *Software quality: Future perspectives on software engineering quality: 13th international conference, SWQD 2021, vienna, austria, january 19–21, 2021, proceedings 13*, 94–105. Springer.

Minderer, Matthias, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, och Mario Lucic. 2021. ”Revisiting the calibration of modern neural networks”. *Advances in Neural Information Processing Systems* 34: 15682–94.

Minsker, Stanislav, Sanvesh Srivastava, Lizhen Lin, och David B Dunson. 2017. ”Robust and scalable Bayes via a median of subset posterior measures”. *The Journal of Machine Learning Research* 18 (1): 4488–4527.

Murshed, MG Sarwar, Christopher Murphy, Daqing Hou, Nazar Khan, Ganesh Ananthanarayanan, och Faraz Hussain. 2021. ”Machine learning at the network edge: A survey”. *ACM Computing Surveys (CSUR)* 54 (8): 1–37.

Mäkinen, Sasu, Henrik Skogström, Eero Laaksonen, och Tommi Mikkonen. 2021. ”Who needs MLOps: What data scientists seek to accomplish and how can MLOps help?” I *2021 IEEE/ACM 1st workshop on AI engineering-software engineering for AI (WAIN)*, 109–12. IEEE.

Mökander, Jakob, Maria Axente, Federico Casolari, och Luciano Floridi. 2022. ”Conformity assessments and post-market monitoring: A guide to the role of auditing in the proposed European AI regulation”. *Minds and Machines* 32 (2): 241–68.

Nguyen, Giang, Stefan Dlugolinsky, Martin Bobák, Viet Tran, Álvaro López García, Ignacio Heredia, Peter Malík, och Ladislav Hluchỳ. 2019. ”Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey”. *Artificial Intelligence Review* 52 (1): 77–124.

Nielsen, Ian E., Dimah Dera, Ghulam Rasool, Ravi P. Ramachandran, och Nidhal Carla Bouaynaya. 2022. ”Robust Explainability: A tutorial on gradient-based attribution methods for deep neural networks”. *IEEE Signal Processing Magazine* 39 (4): 73–84. https://doi.org/10.1109/MSP.2022.3142719.

Oberkampf, William L, Jon C Helton, Cliff A Joslyn, Steven F Wojtkiewicz, och Scott Ferson. 2004. ”Challenge problems: uncertainty in system response given uncertain parameters”. *Reliability Engineering & System Safety* 85 (1–3): 11–19.

Ong, Sylvie C. W., Shao Wei Png, David Hsu, och Wee Sun Lee. 2010. ”Planning under uncertainty for robotic tasks with mixed observability”. *The International Journal of Robotics Research* 29 (8): 1053–68. https://doi.org/10.1177/0278364910369861.

OpenAI. 2023. ”GPT-4 System Card”. https://cdn.openai.com/papers/gpt-4-system-card.pdf.

Opricovic, Serafim, och Gwo-Hshiung Tzeng. 2007. ”Extended VIKOR method in comparison with outranking methods”. *European Journal of Operational Research* 178 (2): 514–29. https://doi.org/10.1016/j.ejor.2006.01.020.

Paleyes, Andrei, Raoul-Gabriel Urma, och Neil D Lawrence. 2022. ”Challenges in deploying machine learning: a survey of case studies”. *ACM Computing Surveys* 55 (6): 1–29.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, m.fl. 2019. ”PyTorch: An imperative style, high-performance deep learning library”. *Advances in neural information processing systems* 32.

Piironen, Juho, och Aki Vehtari. 2017. ”Sparsity information and regularization in the horseshoe and other shrinkage priors”. *Electronic Journal of Statistics* 11 (2): 5018–51.

Pleiss, Geoff, och John P Cunningham. 2021. ”The limitations of large width in neural networks: A deep Gaussian process perspective”. *Advances in Neural Information Processing Systems* 34: 3349–63.

Ras, Gabrielle, Ning Xie, Marcel van Gerven, och Derek Doran. 2022. ”Explainable Deep Learning: A Field Guide for the Uninitiated”. *Journal of Artificial Intelligence Research* 73 (januari): 329–96. https://doi.org/10.1613/jair.1.13200.

Rasmussen, Carl Edward. 2004. ”Gaussian processes in machine learning”. I *Advanced lectures on machine learning: ML summer schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, revised lectures*, redigerad av Olivier Bousquet, Ulrike von Luxburg, och Gunnar Rätsch, 63–71. Berlin, Heidelberg: Springer Berlin Heidelberg.

Ray, Kolyan, Botond Szabó, och Gabriel Clara. 2020. ”Spike and slab variational Bayes for high dimensional logistic regression”. *Advances in Neural Information Processing Systems* 33: 14423–34.

Ren, Pengzhen, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, och Xin Wang. 2021. ”A comprehensive survey of neural architecture search: Challenges and solutions”. *ACM Computing Surveys (CSUR)* 54 (4): 1–34.

Ribeiro, Marco Tulio, Sameer Singh, och Carlos Guestrin. 2016. ”’Why Should I Trust You?’: Explaining the Predictions of Any Classifier”. I *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–44. KDD ’16. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778.

Rudin, Cynthia. 2019. ”Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead”. *Nature Machine Intelligence* 1 (5): 206–15. https://doi.org/10.1038/s42256-019-0048-x.

Sagi, Omer, och Lior Rokach. 2018. ”Ensemble learning: A survey”. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8 (4): e1249.

Sambasivan, Nithya, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, och Lora M Aroyo. 2021. ”“Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI”. I *proceedings of the 2021 CHI conference on human factors in computing systems*, 1–15.

Sculley, David, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, och Dan Dennison. 2015. ”Hidden technical debt in machine learning systems”. *Advances in neural information processing systems* 28.

Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, och Dhruv Batra. 2017. ”Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization”. I , 618–26. https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html.

Sensoy, Murat, Lance Kaplan, och Melih Kandemir. 2018. ”Evidential deep learning to quantify classification uncertainty”. *Advances in neural information processing systems* 31.

Serré, Lynne, Maude Amyot-Bourgeois, och Brittany Astles. 2021. ”Use of Shapley Additive Explanations in Interpreting Agent-Based Simulations of Military Operational Scenarios”. I *2021 Annual Modeling and Simulation Conference (ANNSIM)*, 1–12. https://doi.org/10.23919/ANNSIM52504.2021.9552151.

Shen, Kao-Yi, och Gwo-Hshiung Tzeng. 2016. ”Contextual Improvement Planning by Fuzzy-Rough Machine Learning: A Novel Bipolar Approach for Business Analytics”. *International Journal of Fuzzy Systems* 18 (6): 940–55. https://doi.org/10.1007/s40815-016-0215-8.

Speith, Timo. 2022. ”A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods”. I *2022 ACM Conference on Fairness, Accountability, and*

Transparency, 2239–50. Seoul Republic of Korea: ACM. https://doi.org/10.1145/3531146.3534639.

Srivastava, Sanvesh, Cheng Li, och David B Dunson. 2018. ”Scalable Bayes via barycenter in Wasserstein space”. *The Journal of Machine Learning Research* 19 (1): 312–46.

Sunberg, Zachary, och Mykel Kochenderfer. 2018. ”Online Algorithms for POMDPs with Continuous State, Action, and Observation Spaces”. arXiv. http://arxiv.org/abs/1709.06196.

Sutton, Richard S., och Andrew G. Barto. 2018. *Reinforcement Learning, Second Edition: An Introduction*. Second edition. Cambridge, Massachusetts London, England: Bradford Books.

Symeonidis, Georgios, Evangelos Nerantzis, Apostolos Kazakis, och George A Papakostas. 2022. ”MLOps-definitions, tools and challenges”. I *2022 IEEE 12th annual computing and communication workshop and conference (CCWC)*, 0453–60. IEEE.

Talts, Sean, Michael Betancourt, Daniel Simpson, Aki Vehtari, och Andrew Gelman. 2018. ”Validating Bayesian inference algorithms with simulation-based calibration”. *arXiv preprint arXiv:1804.06788*.

Teh, Yee, Michael Jordan, Matthew Beal, och David Blei. 2004. ”Sharing clusters among related groups: Hierarchical Dirichlet processes”. *Advances in neural information processing systems* 17.

Thong, Nguyen Tho, Luu Quoc Dat, Le Hoang Son, Nguyen Dinh Hoa, Mumtaz Ali, och Florentin Smarandache. 2019. ”Dynamic interval valued neutrosophic set: Modeling decision making in dynamic environments”. *Computers in Industry* 108: 45–52. https://doi.org/10.1016/j.compind.2019.02.009.

Thong, Nguyen Tho, Florentin Smarandache, Nguyen Dinh Hoa, Le Hoang Son, Luong Thi Hong Lan, Cu Nguyen Giap, Dao The Son, och Hoang Viet Long. 2020. ”A Novel Dynamic Multi-Criteria Decision Making Method Based on Generalized Dynamic Interval-Valued Neutrosophic Set”. *Symmetry* 12 (4). https://doi.org/10.3390/sym12040618.

Tran, Ba-Hien, Simone Rossi, Dimitrios Milios, och Maurizio Filippone. 2022. ”All you need is a good functional prior for Bayesian deep learning”. *Journal of Machine Learning Research* 23 (74): 1–56.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, och Illia Polosukhin. 2017. ”Attention is all you need”. *Advances in neural information processing systems* 30.

Vehtari, Aki, Jonah Gabry, Mans Magnusson, Yuling Yao, Paul-Christian Bürkner, Topi Paananen, och Andrew Gelman. 2022. ”loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models”. https://mc-stan.org/loo/.

Vehtari, Aki, Andrew Gelman, och Jonah Gabry. 2017. ”Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC”. *Statistics and computing* 27 (5): 1413–32.

Von Rueden, Laura, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, m.fl. 2021. ”Informed Machine Learning–A taxonomy and survey of integrating prior knowledge into learning systems”. *IEEE Transactions on Knowledge and Data Engineering* 35 (1): 614–33.

Walley, Peter. 2000. ”Towards a unified theory of imprecise probability”. *International Journal of Approximate Reasoning* 24 (2–3): 125–48.

Wang, Hao, och Dit-Yan Yeung. 2020. ”A survey on Bayesian deep learning”. *ACM Computing Surveys (CSUR)* 53 (5): 1–37.

Wilson, Andrew G, och Pavel Izmailov. 2020. ”Bayesian deep learning and a probabilistic perspective of generalization”. *Advances in neural information processing systems* 33: 4697–4708.

Xu, Wei, Marvin J Dainoff, Liezhong Ge, och Zaifeng Gao. 2023. ”Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI”. *International Journal of Human–Computer Interaction* 39 (3): 494–518.

Yalcin, Ahmet Selcuk, Huseyin Selcuk Kilic, och Dursun Delen. 2022. ”The Use of Multi-Criteria Decision-Making Methods in Business Analytics: A Comprehensive Literature Review”. *Technological Forecasting and Social Change* 174 (januari): 121193. https://doi.org/10.1016/j.techfore.2021.121193.

Yang, Chen, Peng Liang, Liming Fu, och Zengyang Li. 2021. ”Self-claimed assumptions in deep learning frameworks: An exploratory study”. I *Evaluation and assessment in software engineering*, 139–48.

Yao, Yuling, Vehtari, Aki, och Gelman, Andrew. 2022. ”Stacking for non-mixing Bayesian computations: The curse and blessing of multimodal posteriors”. *Journal of Machine Learning Research* 23 (79): 1–45.

Yao, Yuling, Aki Vehtari, Daniel Simpson, och Andrew Gelman. 2018. ”Yes, but did it work?: Evaluating variational inference”. I *International conference on machine learning*, 5581–90.

Yazdani, Morteza, Pascale Zarate, Edmundas Kazimieras Zavadskas, och Zenonas Turskis. 2019. ”A Combined Compromise Solution (CoCoSo) Method for Multi-Criteria Decision-Making Problems”. *Management Decision* 57 (9): 2501–19. https://doi.org/10.1108/MD-05-2017-0458.

Ying, Zhitao, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, och Jure Leskovec. 2019. ”GNNExplainer: Generating Explanations for Graph Neural Networks”. I *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/hash/d80b7040b773199015de6d3b4293c8ff-Abstract.html.

FOI är en huvudsakligen uppdragsfinansierad myndighet under Försvarsdepartementet. Kärnverksamheten är forskning, metod- och teknikutveckling till nytta för försvar och säkerhet. Organisationen har cirka 1000 anställda varav ungefär 800 är forskare. Detta gör organisationen till Sveriges största forskningsinstitut. FOI ger kunderna tillgång till ledande expertis inom ett stort antal tillämpningsområden såsom säkerhetspolitiska studier och analyser inom försvar och säkerhet, bedömning av olika typer av hot, system för ledning och hantering av kriser, skydd mot och hantering av farliga ämnen, IT-säkerhet och nya sensorers möjligheter.

FOI
Totalförsvarets forskningsinstitut
164 90 Stockholm

Tel: 08-55 50 30 00
Fax: 08-55 50 31 00

www.foi.se