Utredning av stödverktyg och metodik för utvärdering av AI-metoder

SIDNEY RYDSTRÖM OCH RONNIE JOHANSSON

FOI-R--5453--SE
ISSN 1650-1942 Februari 2023

Sidney Rydström och Ronnie Johansson

Utredning av stödverktyg och metodik för utvärdering av AI-metoder

Titel	Utredning av stödverktyg och metodik för utvärdering av AI-metoder
Title	Investigation of support tools and methodology for evaluation of AI methods
Rapportnr/Report no	FOI-R--5453--SE
Månad/Month	Februari
Utgivningsår/Year	2023
Antal sidor/Pages	42
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/eamesBot

Sammanfattning

Utveckling av metoder avseende *artificiell intelligens* (AI) sker i imponerande fart vilket medför omfattande krav på effektiv utredning gällande möjlig nytta och tillämpning. Utredningen presenterad i den här rapporten baseras på utvärdering av AI-metoder genom tillämpning av verktyg för jämförelse, prestandamätning och demonstration.

Området för hur man ska jämföra och utvärdera AI-system bedöms vara stort och sakna ett universellt verktyg varför det i dagsläget krävs domänspecifika resurser. Emellertid har nyttiga beröringspunkter identifierats så som *reproducerbarhet* genom exempelvis versionshantering, katalogisering av experiment, och strömlinjeformad dokumentering. För att matcha områdets imponerande utvecklingstakt bedöms det avgörande att tillämpa verktyg med god flexibilitet för användarna.

Riktmärkning används vanligen som benämning för utvärdering av prestanda genom jämförelser. Utöver en generell beskrivning av användning av riktmärkning inom AI, och delmängden maskininlärning (ML) specifikt, redogör rapporten för verktyg avsedda för framtagning och sammanställning av resultat, samt vidare jämförelser. Riktmärkning baseras vanligen på datamängder vilket medför begränsningar varför *omgivningsbaserad utvärdering* är kartlagd för exempelvis utvärdering av förstärkt inlärning och multiagentsystem.

Inom industrin tillämpas MLOps som metodpraxis för utveckling, distribution och produktionssättning av ML-modeller, vanligen genom teknikinfrastruktur benämnd AI/ML-systemplattformar med varierande omfattningsgrader. Vissa delar av systemplattformarna bedöms relevanta men alternativen är många varför verktyg för att kartlägga och jämföra tillgängliga plattformar presenteras.

Nyckelord: Artificiell intelligens, AI, maskininlärning, utvärdering, riktmärkning, reproducerbarhet, stödverktyg

Summary

Development of methods regarding *artificial intelligence* (AI) takes place at an impressive speed, which entails extensive demands for effective investigation regarding possible benefits and applications. This report concerns the evaluation of AI methods through tools for comparisons, performance measurement, and demonstration.

The topic of how to compare and evaluate AI systems is extensive and lacks a universal tool, which is why domain-specific resources are currently required. However, the report identifies useful common denominators, such as *reproducibility* through, for example, version control, cataloguing of experiments, and streamlined documentation. Considering the area's impressive rate of development, it is crucial to apply tools with good flexibility for users.

Benchmarking is usually the term used for evaluating performance through comparisons. The report describes the usage within AI and its subfield machine learning (ML) specifically, apart from tools intended for producing, compiling, and comparing the results. Usually, benchmarking uses data sets, which entails limitations; hence, *environment-based evaluation* is required, for example, evaluation of reinforcement learning and multi-agent systems.

In industry, MLOps is a methodology for developing, distributing and deploying ML models, usually through technology infrastructures called *AI/ML platforms*. Some parts of the platforms are relevant; however, there are many available alternatives hence the presentation of services for investigating and comparing the platforms.

Keywords: Artificial intelligence, AI, machine learning, evaluation, benchmarking, reproducibility, support tools

Innehållsförteckning

1 Introduktion

Rapporten redovisar en utredning av stödverktyg och metodik för utvärdering av metoder tillhörande artificiell intelligens (AI). Under de senaste åren har intresset ökat för AI-metoder, särskilt baserat på djupinlärning (DL), och utvecklingen sker i rask takt gällande nya metoder, algoritmer och implementationer. Omfattningen är imponerande men överväldigande, varför det är tidsödande och kostsamt att utreda vilka landvinningar som är lämpliga och mogna för olika tillämpningar. Stödverktyg för utvärdering av AI-metoder kan utgöra ett sätt att avgöra om en viss metod är lämplig för ett visst problem.

1.1 Bakgrund

Testningsförfarandet för utveckling av traditionella mjukvarusystem är relativt välkänt och kartlagt, men mjukvarusystem baserade på AI ställer nya frågor. AI-system är ofta komplexa och inte sällan icke-deterministiska, varför de kräver särskild hantering [1].

Det finns ett antal olika sätt att utvärdera mjukvarusystem. På systemnivå kan man analysera systemets *pålitlighet* (eng. *dependability*), en egenskap som vidare bryts ner i *tillgänglighet* (eng. *availability*), *tillförlitlighet* (eng. *reliability*), och *underhållbarhet* (eng. *maintainability*) [2]. Traditionellt utvärderas algoritmer i termer av dess *tidskomplexitet*[1] (eng. *time complexity*) och *minneskomplexitet*[2] (eng. *space*) [3]. Ibland tillskrivs algoritmer vissa kvaliteter som exempelvis den speciella egenskapen *anytime* [4] som gäller för de tidskrävande algoritmer som inkrementellt beräknar allt bättre lösningar (exempelvis optimeringsalgoritmer) och som därmed när som helst kan returnera ett svar.

Det finns även sätt att utvärdera som är specifika för AI-system. AI-system är ofta specialiserade för en viss tillämpning (exempelvis för att spela schack eller rekommendera produkter) och kan inte användas för generell problemlösning, en egenskap som förknippas med mänsklig intelligens. Det välkända Turingtestet (även kallat ”imitation game”, ”härmningsleken”) från 1950 [5] var det första försöket att beskriva ett test av artificiell generell intelligens (AGI). Turingtestet har i dagsläget spelat ut sin roll som mått på AGI då det både har kritiserats för att vara subjektivt, inte tillräckligt väl fånga den mänskliga intelligensens olika uttryck och för att vissa chattbotar nu är så bra att testet i princip är avklarat. Nya

[1] Algoritmer med hög (exempelvis exponentiell) tidskomplexitet tenderar att enbart vara tillämpliga på små problem.

[2] I datorteknikens barndom var minneskomplexiteten (i arbetsminne och lagring) en stor utmaning. När minneskretsar och hårddiskar ökade i storlek samtidigt som de minskade i pris blev detta utvärderingskriterium mindre viktigt, för att sedan öka i vikt igen i och med de senaste årtiondenas stora tillgång på data och omfattande maskininlärningsmodeller.

praktiska test har på senare år föreslagits [6], men även dessa har karaktären av att de inriktas på någon viss aspekt av AGI, och att de kan kritiseras för att vara för svepande och känsliga för manipulation.

AI-området kan delas in i olika mer eller mindre separata delområden exempelvis maskininlärning (ML), språkteknologi, datorseende, expertsystem, intelligenta agenter, evolutionära metoder, sökning och planering. Utvecklingen inom ML det senaste decenniet har dock tenderat att sudda ut gränserna mellan delområdena, men delområdena har olika syften och kräver därmed olika utvärderingsmått. *Prediktiv modellering* (exempelvis klassificering) och *deskriptiv modellering* (exempelvis klustring) är olika delområden inom ML och i det förstnämnda fallet används prestandamått som *träffsäkerhet* (eng. *accuracy*), *precision*, *minne* (eng. *recall*) [3], *F1-score*, och i det sistnämnda fallet exempelvis *silhouette score*[4].

Inom flertalet discipliner tillämpas konceptet *testbädd*, vilket inbegriper ett ramverk för att utföra testning av idéer och prototyper [7]. AI är, som nämnts, ett omfattande samlingsbegrepp där exempelvis delmängden ML bland annat innefattar *övervakad inlärning*, *oövervakad inlärning* och *förstärkt inlärning* (eng. *reinforcement learning*, RL), vilka alla ställer olika krav på ett visst utvärderingsverktyg [8]. Inom industrin används begreppet MLOps som beteckning för metodpraxis för att exempelvis utveckla och testa ML-modeller, vanligen med stöd av teknikinfrastruktur benämnd *AI/ML-systemplattformar* [9].

Utvärdering av prestanda kan även genomföras genom jämförelser, vilket frekvent benämns som *riktmärkning* [10]. För AI kan genomförandet av jämförelser exempelvis inriktas på tävlingar mellan olika metoder, eller testning mot mänsklig förmåga för uppgiften [11]. Vanligen består konceptet riktmärkning inom ML primärt av två komponenter: datamängd och utvärderingsmått [12]. För utvärdering av viss typ av uppgifter, exempelvis RL, krävs representation som inte ryms i traditionella datamängder varför o*mgivningsbaserad utvärdering* blir aktuellt [13].

1.2 Syfte

Syftet med arbetet är att utreda möjligheten att utvärdera AI-metoder med någon form av verktyg för jämförelse, prestandamätning och demonstration, med stöd av exempelvis riktmärkesdatamängder.

[3] https://en.wikipedia.org/wiki/Precision_and_recall (besökt 2023-02-17)

[4] https://en.wikipedia.org/wiki/Silhouette_(clustering) (besökt 2023-02-17)

1.3 Avgränsningar

Rapporten tar inget ställningstagande kring eventuell problematisering avseende aspekter som öppen källkod, kommersiella resurser, licenser, ekonomiska kostnader och tillgång till beräkningsresurser. Fokus i utredningen ligger på metoder för riktmärkning, omgivningsbaserad utvärdering och systemplattformar.

1.4 Läsanvisningar

Kapiteldispositionen baseras i huvudsak på indelning av stödverktyg enligt generell metodik för utvärdering av AI-metoder, varför visst överlapp förekommer.

Kapitel 2 avhandlar aspekter av reproducerbarhet vilket är centralt för tillförlitlighet och för att säkerställa rättvis utvärdering av system.

Kapitel 3 avhandlar riktmärkning för AI och särskilt ML. Processen är jämförelse- och utvärderingsorienterad varför verktyg för detta presenteras.

Kapitel 4 avhandlar omgivningsbaserad utvärdering vilket baseras på olika interaktiva miljöer för lärande, testning och jämförelser.

Kapitel 5 avhandlar utvecklingsorienterade verktyg benämnda som AI/ML-systemplattformar som inbegriper konceptet testbädd.

Kapitel 6 avhandlar diverse uppslag som uppdagats under arbetets gång och bedöms relevanta, men som med respekt för arbetets omfattning enbart studerats begränsat, eller saknar naturlig insortering i tillämpad disposition.

Kapitel 7 avhandlar författarnas rekommendationer utifrån de aspekter och resurser de föregående kapitlen identifierat.

Bilaga A består av en förteckning för de centrala förkortningar som förekommer i rapporten.

2 Reproducerbarhet

Inom mjukvaruutveckling är versionshantering av kod redan praxis. En trend inom ML är inkorporering av tjänster för att versionshantera från start till mål, så kallad *end-to-end support*. ML beskrivs ofta innefatta tre essentiella delar: data, kod och modeller, varpå diverse verktyg försöker tillgodose respektive behov, som del- eller helhetslösningar [14]. Reproducerbarhet utgör en ultimat standard gentemot vilken vetenskapliga påståenden utvärderas. Resultat som inte går att återskapa betraktas således inte som vetenskapliga upptäckter [15]. För att uppnå reproducerbarhet av ML-system krävs detaljerad dokumentering då små avvikelser (till exempel såddvärde) kan ge stora skillnader i resultat [16].

De inledande avsnitten 2.1 och 2.2 fokuserar på tekniska lösningar medan avsnitt 2.3 redogör för dokumentering i textformat.

2.1 Docker

Docker[5] är ett projekt med öppen källkod med ändamålet att automatisera distributionen av program genom självförsörjande containrar. Om än besläktat, ska det inte förväxlas med företaget *Docker Inc*[6] som utvecklar verktyg runt Docker. Med container avses en avbildning av en mjukvarumiljö inklusive beroenden och konfigurationer. Utöver portabilitet uppnås även kontroll över arbetsflödet för mjukvarans livscykel. Användandet av Docker är utbrett till den grad att det närmast utgör en standard, och förväntas framöver användas i alla datacenter: molnbaserade såväl som i egna lokaler [17].

2.2 Kubernetes

Kubernetes[7] (även känt som K8s) är ett öppet källkodsprojekt som används för att automatisera distribution, skalning och hantering av containerbaserade applikationer, till exempel Docker. Kubernetes är ett verktyg som orkestrerar, åt användaren, virtuella maskiner till ett gemensamt kluster för att tillgodose ett behov av att skala ut ett program. Tjänsten är även kompatibel med att köra på en enskild maskin som en utvecklingsmiljö [17].

[5] https://github.com/moby/moby (besökt december 2022)

[6] https://www.docker.com/ (besökt december 2022)

[7] https://kubernetes.io/ (besökt december 2022)

2.3 Dokumentering

Data är en av hörnstenarna för ML, men trots det saknas en enhetlig standard för dokumentering av datamängder [18]. Som en konsekvens innehåller datamängder ofta fel, är bristfälligt dokumenterade med oklara framställningsprocesser, samt underhållet är bristfälligt. Utöver att en datamängds livscykel bör innefatta dokumentation, bör den även granskas, revideras, och underhållas [19].

Som ett konkret förslag, med inspiration från elektronikindustrin där varje komponent levereras med tillhörande beskrivning av dess egenskaper, föreslås *Datasheets for datasets*. Genom att dokumentera och förmedla egenskaperna för datamängder skapas en grund för ökad transparens och reproducerbarhet. Under rubrikerna motivering, komposition, insamlingsprocess, bearbetningsprocess, tillämpningar, tillgängliggörande, och underhåll, ryms totalt 57 framarbetade aspekter [18].

Tidigare lanserade projekt för att uppnå en enhetlig standard har fått begränsad genomslagskraft. *Data Cards* är ett nytt försök till att etablera en standard för vad som ska dokumenteras samt på vilket sätt. Målet med Data Cards är att det ska vara flexibelt, modulärt, utvecklingsbart, tillgängligt och domänagnostiskt [20].

Liknande förslag har presenterats för att strömlinjeforma modelldokumentation genom *Model Cards*. Sådan dokumentation ska till exempel erbjuda insikter kring träningsförfarandet, modellarkitektur, lämpliga tillämpningar och begränsningar med modellen. Det möjliggör en jämförelse av modeller bortom utvärderingsmått om än utgör ett ben i konceptet [21].

3 Riktmärkning

Som introducerats i avsnitt 1.1 inbegriper riktmärkning flera aspekter. Användningen av riktmärkning inom AI beskrivs i avsnitt 3.1, medan avsnitt 3.2 fokuserar på ML. Avsnitt 3.3 beskriver diverse verktyg för att sammanställa resultat för jämförelser. Avsnitt 3.4 redogör för resurser som fokuserar på att tillhandahålla data.

3.1 Riktmärkning inom AI

Att använda tester framtagna för människor som utvärdering av AI kan vara missvisande. För det första kan AI lära sig saker som människor har brister gällande. För det andra antar de grundläggande beståndsdelar av mänsklig intelligens vilket uppgiftsinriktade (specialiserade) AI-system sällan inbegriper, till exempel är AI-system ofta begränsade till en viss domän och saknar därtill ofta förmågan att avgöra om ett problem ligger utanför dess kompetens. Därför krävs goda kunskaper om människan och AI för att designa tester just för det ändamålet och på så sätt möjliggöra rättvisande jämförelser [22].

Utvärdering av AI-system särskiljer sig från många andra mjukvaru- och hårdvarusystem, vilket härleds till avsaknaden av formell beskrivning av hur problemet ska lösas. Utvecklingen av teknik går allt snabbare och systemen blir allt mer sofistikerade, vilket försvårar den redan komplicerade uppgiften att formulera och designa meningsfulla och tillförlitliga AI-test. Det förespråkas att utvärderingen ska bli mer iterativ och utmanande, i syfte att göra utvecklingen mer verklighetsförankrad [22].

Det finns en uppsjö av riktmärken och tävlingar för att empiriskt jämföra och utvärdera AI-system. Däremot saknas en enhetlig katalogisering av resurserna. Att finna bra och lämpliga riktmärken för att utvärdera ett AI-system är därför ett svårt problem, särskilt i kontrast till att utvärdera mänsklig intelligens som har strömlinjeformade test. Liksom riktmärken erbjuder tävlingar ett relativt objektivt sätt att mäta utveckling inom AI. Ofta utgörs momentet av en begränsad del av ett större problem och framsteg bör värderas med det i åtanke [22].

Stanford University[8] släpper årligen *The Artificial Intelligence Index Report*. I utgåvan 2021 avhandlar ett kapitel teknisk prestanda. Där redogörs för den aktuella förmågan av prestanda inom diverse domäner med stark utgångspunkt från diverse datamängder för riktmärkning. Resultaten redovisas över tid vilket möjliggör en beskrivning av historisk utveckling [11].

[8] https://www.stanford.edu/ (besökt december 2022)

3.2 Riktmärkning inom ML

Utvecklingen inom ML har varit imponerande, och en betydelsefull roll tillskrivs de datamängder för riktmärkning som har drivit utvecklingen av modellarkitekturer genom att tillhandahålla en gemensam grund för att utvärdera modellprestanda [23]. Emellertid ger metodiken upphov till svagheter och sårbarheter, då bland annat flera erkända datamängder, som används för att utvärdera *State of the art*-metoder, innehåller felannoteringar[9]. Konsekvensen är att utvecklingen kan styras av att särskilja felaktiga specialfall snarare än att lösa den större uppgift riktmärket är surrogat för [24]. Ett annat problem är att utvecklingen på riktmärken mättas allt snabbare inom flera uppgifter. Som en konsekvens höjs vissa röster gällande ett skifte från modellcentrerad ML till datacentrerad ML där datamängderna är föränderliga med tesen att det bättre återspeglar verkligheten [23].

Vikten av att välja rätt riktmärke med avseende på den verkliga tillämpningen går inte att förringa. Den tillhörande datamängden definierar världen för i vilken det är möjligt att extrahera information för den lärande algoritmen. Om datamängden inte är representativ för det verkliga fallet förringas förmågan att generalisera för modellen. Därtill fokuserar dagens utvärderingsmått på att bedöma modellens lämplighet i relation till den betraktade datamängden. Därav kvantifieras inte hur väl datamängden representerar den verkliga tillämpningen och dess kravställning [25].

Riktmärkesuppgifter organiserar och koordinerar ML-forskning genom att konkret bistå med mätbarhet gentemot ett gemensamt mål. En kartläggning av riktmärken under tidsperioden 2015 – 2020 konstaterar att allt färre riktmärken används inom respektive uppgiftsdomän, samt att de dominerande datamängderna är framtagna av en handfull institutioner. Det finns tendenser att datamängder för riktmärkning utvecklade för en viss uppgiftsdomän tillämpas för en annan uppgift, trots tillgång på uppgiftsspecifika datamängder. Faktum är att det finns en positiv trend för skapande av nya datamängder för riktmärkning, men erkännandet uteblir som en konsekvens av bristen på faktisk tillämpning [12].

Neural Information Processing Systems[10] (NeurIPS) är en årlig konferens inriktad på ML. Med start 2021 lanserade de en särskild gren för datamängder och riktmärkning, i syfte att öka publiceringar och diskussioner kring de aspekterna av ML, som historiskt sett ägnats lite utrymme gällande accepterade bidrag [26]. För 2021 års konferens tillhandahålls en förteckning[11] av de 174 accepterade bidragen.

---

[9] Med annotering avses den beskrivande klass som data tilldelats. För felannoteringar, se exempelvis https://labelerrors.com/ (besökt december 2022)

[10] https://neurips.cc/ (besökt december 2022)

[11] https://nips.cc/Conferences/2021/DatasetsBenchmarks/AcceptedPapers (besökt december 2022)

3.3 Jämförelseplattformar

För att utvärdera resultat av riktmärken och till viss del tävlingar används lämpligen en plattform för att sammanställa resultaten och därigenom möjliggöra jämförelser. För utvärdering av AI-metoder erbjuder det utöver kartläggning av utveckling inom området även en möjlig inspiration för hur jämförelser kan utföras.

3.3.1 MLCommons

MLCommons[12] är ett ingenjörsvetenskapligt konsortium bestående av parter från industrin och akademin. Målet med verksamheten är att accelerera innovation för ML samt dess positiva inverkan på samhället. Verksamheten härrör från framtagning av ett riktmärke och tillhandahåller idag flertalet varav huvudsakligen två redovisas nedan.

3.3.1.1 MLPerf

MLPerf[13] består av två riktmärken med syfte att driva teknologisk utveckling, och mäta resultat över tid, främst för hårdvara men även för mjukvara. Verktyget ämnas underhållas samt vidareutvecklas. Detta för att minska risken att utvecklas efter ett riktmärke snarare än en verklig tillämpning. De två komponenterna är [27]:

- *MLPerf Training* som utgår från att träna en modell på en given datamängd tills ett specifikt värde för ett förutbestämt utvärderingsmått (till exempel träffsäkerhet) uppnås. Jämförelser sker på två nivåer: en stängd nivå där modellstrukturen är förutbestämd av riktmärkesuppgiften, och en öppen nivå där valfri modell appliceras. Syftet med MLPerf Training är att tillhandahålla en plattform för att jämföra hårdvaruprestanda, men även främja innovativ mjukvaruutveckling. Prestandan för träning mäts och jämförs främst genom mängden data processad per sekund och total tid. Variansen för utvärderingsmåttet är emellertid en betydelsefull komponent varför experimenten upprepas och sammanvägs för jämförelse [27].
- *MLPerf Inference* som består av att en färdigtränad modell bearbetar en serie av indata givet ett av fyra scenarier. Utvärderingen av respektive scenario är förknippat med visst utvärderingsmått. Likt för MLPerf Training tillhandahålls en stängd och en öppen nivå [27].

[12] https://mlcommons.org/ (besökt december 2022)

[13] Exempelvis https://mlcommons.org/en/news/mlperf-training-4q2022/ (besökt januari 2023)

3.3.1.2 Dynabench

Dynabench[14] är ursprungligen en plattform för att skapa dynamiska datamängder och jämföra prestanda för modeller gällande datorhantering av språkteknologi. Syftet är att överbrygga problematiken kring att modeller uppvisar förträffliga resultat på riktmärkesuppgifter, samtidigt som de fallerar gällande enkla verkliga utmaningar. Dilemmat tillskrivs förenklingar och begränsningar avseende den ultimata uppgiften, varpå en lösning föreslås vara att dynamiskt utvärdera modeller mot människor. Tillvägagångssättet kan beskrivas som en iterativ process mellan skapande av ny data av människor och utvärdering av modeller [28]. Principen liknar en lösning som tidigare föreslagits av [29] för generell utvärdering av AI-system, bestående av två oberoende parter där den ena parten föreslår modeller och den andra parten utvärderar modellerna.

Sedan lanseringen har flera tillägg inkorporerats. *Dynascore* lanserades som en kontrast till ett enskilt utvärderingsmått, till exempel träffsäkerhet. Likt föreslaget av [30] tillhandahålls jämförelse baserat på nytta, inspirerat av teori från mikroekonomi. Användaren tillåts sammanväga diverse mått med olika vikter för att erhålla en ranking efter dennes preferens. Exempelvis kan beräkningstid, robusthet och träffsäkerhet sammanvägas, och användas som ett gemensamt utvärderingsmått. I de fall som flera datamängder ingår i samma riktmärke, kan användaren även tillskriva dem olika vikter [31].

Vidare har arbete för att tillhandahålla reproducerbarhet och bakåt-/framåtkompabilitet inkorporerats [31], samt att tillåta användare att vara värd för nya AI-uppgifter [32]. Ett nytt tillskott är att plattformen är värd för *DataPerf*[15], riktmärkning för datacentrerad AI. Jämförelser möjliggörs i två olika nivåer: en öppen nivå som tillåter bidrag utan tillhörande kod, och en stängd nivå med krav på tillhörande kod samt förbjuder manuella komponenter i databearbetningen för att främja reproducerbarhet [23].

3.3.2 EvalAI

EvalAI[16] är en plattform för att utvärdera AI- och ML-modeller genom mänsklig inblandning och dynamiska miljöer. Det möjliggör utvärdering av interaktiva agenter eftersom förstärkt inlärning (RL) kräver nya miljöer vilka inte kan representeras med traditionella utvärderingsdatamängder. Plattformen utgår från olika utmaningar representerade av tillhörande topplistor som uppdateras genom att deltagare laddar upp Docker-containrar av färdigtränade modeller som av anordnaren av utmaningen utvärderas på testmiljöer. Jämfört med traditionell riktmärkning, uppges plattformen vara särskilt lämplig för multimodala uppgifter

[14] https://dynabench.org/ (besökt december 2022)

[15] https://dataperf.org/ (besökt december 2022)

[16] https://eval.ai/ (besökt december 2022)

(exempelvis kombination av bild- och textrepresentation), motiverat av den mänskliga inblandningen i utvärderingsprocessen [13].

3.3.3 CodaLab

CodaLab[17] är ett projekt med öppen källkod där tjänsten *CodaLab Competitions* har ändamålet att anordna tävlingar kopplade till datavetenskap, däribland ML. Plattformen möjliggör tävlan genom inlämning av framtaget resultat, eller genom inlämning av kod där plattformen tar fram tävlingsresultaten från inskickad Docker-container. Tävlingarna går att dela upp i flera faser, och baseras på anpassade utvärderingsmått [33].

En ny version av plattformen, *Codabench*[18] är under utveckling och finns tillgänglig som betaversion. Den nya plattformen är särskilt inriktad på riktmärkning över tid, snarare än tidsbegränsade tävlingar [33]. Plattformen har en gratis onlinetjänst, men går även att köra lokalt. Likt sin föregångare baseras tjänsten på öppen källkod varför den tillhandahållna teknikstacken[19] är väldokumenterat tillgängliggjord [34].

3.3.4 BIG-bench

BIG-bench[20] (beyond the imitation game[21] benchmark) är ett riktmärkesverktyg för språkmodeller. Verktyget innehåller för närvarande fler än 200 utmaningar för språkmodeller som bidragits av över 130 institutioner och användarna uppmanas lägga till fler efter hand. Syftet är att få grepp om den snabba utvecklingen av språkmodeller och kvantitativt kunna följa utvecklingen [35]. Utmaningarna omfattar bland annat problem inom matematik, lingvistik, fysik, och fördomar och försöker pressa gränserna för vad nuvarande språkmodeller klarar av.

3.4 Datamängder

Utöver de jämförelseplattformar som nämns i avsnitt 3.3 har projektet resulterat i ett uppslag kring identifierade datamängder. Särskiljningen till detta avsnitt utgörs av mer domänspecifika riktmärkesresurser eller en plattform som primärt inte är inriktad på jämförelser. Presenterade förteckningar ämnar inte vara allom-

[17] https://codalab.org/ (besökt december 2022)

[18] https://www.codabench.org/ (besökt december 2022)

[19] Med stack avses alla beståndsdelar.

[20] https://github.com/google/BIG-bench (besökt februari 2023)

[21] Med en blinkning till Alan Turings imitation-game-test (se avsnitt 1.1)

fattande utan representerar ett urval av tillgängliga publika resurser. Redovisningen sker i uppdelning av resurser som är domänspecifika och de som utgörs av samlingar av resurser.

3.4.1 Domänspecifika datamängder

SuperGLUE[22] är inriktat på förståelse av naturliga språk genom åtta olika uppgifter för att jämföra förmågor för människor och maskiner. SuperGLUE är en uppdatering från föregångaren *GLUE* [36].

Multilingual Spoken Words[23] är en ljuddatamängd som vid lansering omfattade tal från över 23 miljoner ljudinspelningar fördelat på 50 olika språk. För flera språk är detta den första tillgängliga resursen av sitt slag [37].

Audio Set[24] består av videos som har annoterats av människor angående förekomst av olika ljud. Tillämpningsområden inkluderar detektering och klassificering av akustiska händelser [38].

MOTChallenge[25] består av datamängder inriktade på datorseende för multipla objekt. Specifikt inriktar sig resursen på segmentering och målföljning. Sedan lansering har datamängderna utökats löpande [39].

Perception test[26] utgår från faktiska videor inriktade på att utvärdera multimodala modeller avseende ljud, seende och text. Material är insamlat från runt hundra deltagare och består av 11 600 videosekvenser vilka har annoterats enligt exempelvis objektens position och förändringar i sekvensen, samt en beskrivning av händelsen [40].

ParlAI[27] är en plattform som samlar resurser för dialogmodeller. Ambitionen är att tillhandahålla ett enat programmeringsbibliotek för att träna, testa och dela modeller. Processen inbegriper mänsklig interaktion med modellerna och datainsamlingen. Det finns över 100 datamängder tillgängliga, utöver referens- och förtränade modeller [41].

[22] https://super.gluebenchmark.com/ (besökt december 2022)

[23] https://mlcommons.org/en/multilingual-spoken-words/ (besökt december 2022)

[24] https://research.google.com/audioset/ (besökt december 2022)

[25] https://motchallenge.net/ (besökt december 2022)

[26] https://github.com/deepmind/perception_test/ (besökt december 2022)

[27] https://parl.ai/ (besökt december 2022)

3.4.2 Samlingar av datamängder

Utöver resurser inriktade på särskilda uppgifter, finns även tjänster (se Tabell 1) som tillhandahåller en samling publika datamängder där domän, kvalitet, dokumentation och omfattning kan variera. Samlingar av datamängder med olika egenskaper erbjuder en möjlighet för att robusttesta och jämföra modeller. Vissa av resurserna associerar även datamängderna med tillämpade metoder, till exempel *Kaggle* genom användarnas anteckningsböcker, och *Papers with Code* genom hänvisning publikationer i databasen.

Tabell 1: Samlingar av datamängder. Redovisad statistik gör sig gällande för den 9 december 2022.

Namn	**Antal datamängder**
Kaggle[28]	185 119
Hugging Face[29]	15 781
Papers with Code[30]	7 473
OpenML[31]	4 896
UCI Machine Learning Repository[32]	612
Penn Machine Learning Benchmarks[33]	417
Google Research[34]	135

[28] https://www.kaggle.com/ (besökt december 2022)

[29] https://huggingface.co/ (besökt december 2022)

[30] https://paperswithcode.com/ (besökt december 2022)

[31] https://www.openml.org/ (besökt december 2022)

[32] https://archive.ics.uci.edu/, under uppdatering: https://archive-beta.ics.uci.edu/, (besökt december 2022)

[33] https://epistasislab.github.io/pmlb/ (besökt december 2022)

[34] https://research.google/, tillhandahåller även en sökmotor *Dataset Search,* (besökt december 2022)

4 Omgivningsbaserad utvärdering

Omgivningsbaserad utvärdering tillämpas för exempelvis RL och multiagentsystem, där testningsförfarandet inte inbegrips i en datamängd, utan i relation till en simulerad eller verklig omgivning (även kallat miljö). Utvärderingsformen används även vanligen för system som strävar efter att utvärdera artificiell generell intelligens (AGI). Avsnitt 4.1 till avsnitt 4.5 beskriver diverse stödverktyg för relaterad utvärdering.

4.1 Gym

OpenAI[35] *Gym*[36] är ett verktyg för riktmärkning av RL skapat som en resurs för att förena etablerade riktmärken och sätta en standard för nya. Verktyget består av ett API för *Python*[37] som tillhandahåller flera olika miljöer att utvärdera agenter i [42]. Efter stagnerande utveckling och bristande underhåll överlät OpenAI 2021 kontrollen till den ideella organisationen *The Farama Foundation*[38] vars ambition är att samla öppenkällkodsbibliotek för RL. Överlåtandet beskrivs som en långsiktig strategi för att garantera tillgänglighet av exakt samma version av miljöer för att uppnå reproducerbarhet och pålitlighet för forskning inom RL. Numera distribueras det under namnet *Gymnasium*[39] vilket tillhandahåller bakåtkompabilitet till det tidigare biblioteket. Farama tillhandahåller idag åtta bibliotek för RL, däribland *PettingZoo*[40] vilket är motsvarigheten för miljöer med flera agenter [43].

4.2 bsuite

Behaviour Suite (bsuite)[41] är ett Python-bibliotek bestående av en uppsättning experiment för att utvärdera centrala förmågor för RL-agenter. Verktyget ämnar inte utgöra en representation av utvecklingen i stort, utan fokuserar på avgränsade tester för agenternas beteenden för att erhålla jämförelser för grundläggande egenskaper, som exempelvis minnesförmåga [44].

[35] https://openai.com/ (besökt december 2022)

[36] https://www.gymlibrary.dev/ (besökt december 2022)

[37] https://www.python.org/ (besökt december 2022)

[38] https://farama.org/ (besökt december 2022)

[39] https://gymnasium.farama.org/ (besökt december 2022)

[40] https://pettingzoo.farama.org/ (besökt december 2022)

[41] https://github.com/deepmind/bsuite (besökt december 2022)

4.3 Animal AI

Animal AI[42] härrör från en tävling anordnad för NeurIPS och är idag en testbädd bestående av en 3D-simulator byggd med spelverktyget *Unity*[43] med en tillhörande tränings-API för Python. Kognitiva tester av perception och navigation för agenter kan testas genom 900 uppgifter, vilka har inspirerats av djurs förmågor i naturen. Genom testbädden önskas agenternas förmåga att resonera med sunt förnuft kunna utvärderas [45].

4.4 Project Malmo

Project Malmo[44] är en flexibel plattform baserat på datorspelet *Minecraft*[45] vilket erbjuder en komplex tredimensionell värld med gränslös variation av scenarier. Plattformen består av ett abstraktionslager ovanpå datorspelet där API:n har stöd för diverse operativsystem och programmeringsspråk. Om forskningen om AI ska leda till flexibel AI, dvs. AGI, krävs verktyg som tillåter experiment över flera domäner. Project Malmo är tillgängligt som öppenkällkod och har stöd för robotik, datorseende, RL, planering och multiagentsystem, med besläktade områden [46].

4.5 SAGE

Simulator for Autonomy & Generality Evaluation (SAGE) är en plattform för generell maskinintelligens kontrasterat av majoriteten av tillgängliga verktyg som ofta tillåter avgränsade tester. Plattformen är dock fortfarande under utveckling och källkoden för plattformen är planerad att offentliggöras när utvecklingsarbetet är slutfört. Förarbetet beskriver en flexibel kontrollerbar testmiljö och en uppsättning eftersträvansvärda egenskaper [47] baserade på [48]:

1. Determinism
2. Ergodicitet
3. Kontrollerbar kontinuitet
4. Asynkronicitet
5. Dynamism
6. Observerbarhet
7. Kontrollerbarhet
8. Multipla parallella kausala kedjor
9. Flera agenter
10. Periodicitet
11. Reproducerbarhet.

[42] https://github.com/mdcrosby/animal-ai (besökt december 2022)

[43] https://unity.com/ (besökt december 2022)

[44] https://github.com/microsoft/malmo (besökt december 2022)

[45] https://www.minecraft.net/ (besökt december 2022)

4.6 Multiagentsystem

Multiagentsystem består av en uppsättning autonoma entiteter som samarbetar för att lösa en komplex uppgift. Sammansättningen av flera beståndsdelar ställer särskilda krav för utveckling och utvärdering av sådana system. Exempelvis behöver aspekter som koordinering, uppgiftsallokering och säkerhet hanteras. Föreslagna verktyg för att simulera scenarion i utvärderingssyfte är *NetLogo*[46], *Java Agent Development frame work* (JADE)[47], *GAMA*[48] och *Matlab*[49] [49].

Phy-Q är en testbädd och prestandamått för agenter som löser problem i fysiska miljöer [50] (i vilket fall det då typiskt handlar om robotar). Testbädden omfattar 15 stycken typproblem som beskriver fysiska problem som agenter ska försöka lösa. Typproblemen vilka bland annat handlar om rörelse, gravitation, timing, varieras för att utmana agenternas förmåga till generell problemlösning. Syftet med testbädden är att bana väg för agenter med mänsklig förmåga att resonera kring fysiska problem.

[46] http://ccl.northwestern.edu/netlogo/index.shtml (besökt december 2022)

[47] https://jade.tilab.com/ (besökt december 2022)

[48] https://gama-platform.org/ (besökt december 2022)

[49] https://github.com/douthwja01/OpenMAS (besökt februari 2023)

5 AI/ML-systemplattformar

Det finns stora mängder verktyg som inriktas på utveckling av AI/ML-system där utvärderingsaspekten utgör en delmängd. Avsnitt 5.1 beskriver konceptet MLOps och dess utveckling. Avsnitt 5.2 är inriktat på verktyg för att kartlägga tillgängliga resurser. Avsnitt 5.3 redogör för vissa av identifierade resurser.

5.1 MLOps

DevOps utgör metodpraxis för hur mjukvaruapplikationer inom industrin effektivt ska utvecklas, distribueras och sättas i produktion. Arbetssättet utgör grunden för *MLOps* med tillägg av processer för att utveckla ML-modeller, samt inkorporering av de utvecklade modellerna. MLOps ställer krav på flexibilitet för utveckling av lämpliga modeller, samtidigt som modellerna behöver vara kompabila med den infrastruktur som produktionssättningen baseras på. För att möta behoven krävs således en uppsättning tjänster och tillvägagångssätt som möjliggör tillämpning av AI framgångsrikt [9].

Den kommersiella AI-utvecklingen drivs främst av de stora teknikföretagen. Som en konsekvens utkämpas en maktkamp mellan främst *Microsoft*[50], *Amazon*[51] och *IBM*[52] gällande systemplattformar för molntjänster. Utvecklingen beskrivs som plattformisering, ofta grundat på en öppen standard, men där plattformens nytta ökar med tillströmning av användare. Centralt för systemplattformarna är att de bistår med beräkningskraft, men de kan även bidra med att underlätta AI-utveckling. Emellertid uppnår dessa systemplattformar inte kundernas fulla behov vilket resulterar i hybridlösningar [8].

En trend för systemplattformarna är att erbjuda, bland andra lösningar, en tjänst för Automatisk ML (AutoML). Utöver att automatisera processer som hyperparametersökningar inbegriper tjänsten sällan krav på kodningskunskaper utan baseras på ett grafiskt gränssnitt enligt principen dra-och-släpp. Tröskelkraven och tidsåtgången reduceras men så även flexibiliteten gällande valfrihet för användaren [8].

Historiskt sett, tog de ledande teknikföretagen fram sina egna AI/ML-systemplattformar från grunden. Som en demokratiseringsprocess av AI har det skett en explosion av tillgängliga systemplattformar, anpassade för stora som små användare. Utvecklingen har varit överväldigande till den grad att det idag är svårt att

[50] https://www.microsoft.com/ (besökt december 2022)

[51] https://www.amazon.com/ (besökt december 2022)

[52] https://www.ibm.com/ (besökt december 2022)

överblicka och granska vad tjänsterna erbjuder och vad som särskiljer dem från varandra. Att utvärdera tjänster baserat på marknadsföring tenderar att ge upphov till bristfälliga processer. Det försvåras även genom att arkitekturer definieras olika gällande flöden och inkluderade delar [51].

5.2 Undersökande tjänster

AI Infrastructure Alliance[53] (AIIA) skapades för att utgöra en grund för samarbete för företag som utvecklar AI-tjänster. Samarbete kan vara ett sätt att uppnå mer enhetliga tjänster genom utvecklingsprocesser, istället för att etablera tvingande standarder. Kortfattat kan målbilden beskrivas som jakten på ett agnostiskt orkestreringssystem med välfungerande API och väldefinierade kommunikationsstandarder. En liknelse är Kubernetes, fast för ML, dvs. ett verktyg som abstraherar koncept och kommunikation mellan alla lager i en ML-stack [51].

AIIA presenterade för 2022 en rapport gällande AI-infrastruktur, som utöver kartläggningar även redovisar framtidsutsikter. Rapporten rekommenderar i allmänhet företag att inte bygga systemplattformar från grunden, men höjer även ett varningens finger för end-to-end[54] systemplattformar motiverat av områdets ständiga utveckling och att tjänsterna ibland lovar mer än vad de håller. För jämförelser av tjänster har en karta över en ML-stack tagits fram, vilken kan färgläggas efter en tjänsts förmågor varpå visuella jämförelser kan genomföras med andra tjänster. Utöver jämförelser och utvärdering av likvärdiga tjänster är det även användbart i syfte att finna kompletterande tjänster som möter det fulla behovet [52].

Som en ytterligare resurs för att utvärdera landskapet av AI-infrastruktur tillhandahåller AIIA en kartläggning av 80 verktyg graderade i en tabell för ett tjugotal kategorier som t.ex. modellutveckling, finjustering av hyperparametrar, och experimenthantering [53].

Forrester[55] genomförde för det tredje kvartalet 2022 en övergripande kartläggning av AI/ML-systemplattformar där 15 stycken identifierades som framträdande. Bland dessa identifierades *Palantir*[56], *C3 AI*[57] och *DataRobot*[58] som marknadsledare. Systemplattformarna bedömdes enligt 25 kriterier fördelade enligt tre kategorier: nuvarande tjänst, strategi, och marknadsandel [54].

---

[53] https://ai-infrastructure.org/ (besökt december 2022)

[54] End-to-end syftar till ambitionen att hantera alla aspekter av AI/ML livscykeln [52].

[55] https://www.forrester.com/ (besökt december 2022)

[56] https://www.palantir.com/ (besökt december 2022)

[57] https://c3.ai/ (besökt december 2022)

[58] https://www.datarobot.com/ (besökt december 2022)

En del av MLOps överlappar med det här arbetets omfattning, dock förefaller systemplattformar till hög grad fokusera på att tillhandahålla beräkningskapacitet och produktionssättning av utvecklade modeller. Utvärdering av verktyg baserat på företagens hemsidor är tidskrävande och bedöms osäkert. För ingående analyser rekommenderas initial sondering med hjälp av verktyg framtagna av till exempel AIIA, uppföljt av möten med representanter för identifierade resurser.

Generellt bedöms stödet för RL vara relativt begränsat och erbjuds i regel enbart av de stora teknikbolagen. Däremot är graden av stöd generellt svårdefinierat eftersom vissa systemplattformar utgörs av ett enhetligt verktyg medan andra systemplattformar består av en uppsättning av verktyg med varierande grad av integrering. Emellertid blir utvärderingen även vag eftersom systemplattformarna i sin tur beror på annan mjukvara som tillhandahåller vissa metoder för RL, exempelvis *Ray*[59] som har ett bibliotek *RLlib* för att skala upp RL, eller *Tensorflow*[60] som har visst stöd för att träna agenter.

5.3 Exempel på systemplattformar

För att tillgodose användarnas behov så använder systemplattformar frekvent verktyg från tredjepart. Det finns ett återkommande stöd för till exempel *Microsoft Azure*[61] och *Amazon Web Services*[62] (AWS) bland de systemplattformar som inkluderar molnlösningar, ofta med stöd av användning av Docker och Kubernetes. Det dominerande programmeringsspråket förefaller vara Python, ofta med integrerade lösningar för bibliotek som till exempel Tensorflow och *PyTorch*[63].

Tidigare nämnda systemplattformar bedöms fokusera på helhetslösningar, medan den här rapporten är inriktad på den inledande delen av MLOps med till exempel experimenthantering, samarbete och reproducerbarhet. Resterande del av avsnittet avhandlar tre olika systemplattformar med varierande omfattning för tillhandahållna lösningar: *MLflow* fokuserar på att hantera experiment, *Comet* är inriktat på utvärdering av modeller, och *DataRobot* har en tydlig målbild att erbjuda en helomfattande lösning.

[59] https://www.ray.io/ (besökt december 2022)

[60] https://www.tensorflow.org/ (besökt december 2022)

[61] https://azure.microsoft.com/ (besökt december 2022)

[62] https://aws.amazon.com/ (besökt december 2022)

[63] https://pytorch.org/ (besökt december 2022)

5.3.1 MLflow

MLflow[64] är en plattform med öppen källkod ämnad för att hantera ML-livscykeln genom ett generisk API. Ambitionen är att fungera för alla algoritmer, programmeringsspråk (främst stöd för Python) och bibliotek, där gränssnittet tillför värde för användaren genom en strukturerad utvecklingsprocess. Plattformen täcker tre utmaningar: experimentspårning, reproducerbarhet, och modelldriftsättning [55]. Det möjliggörs genom fyra tjänster:

- *MLflow Tracking* är ett verktyg för att dokumentera genomförda experiment. Spårningen inkluderar kod, parametrar, indata, mått, och resultatfiler. Via tillämpning av API eller UI kan experimenten inspekteras och jämföras [55], [56].
- *MLflow Projects* är inriktat på reproducerbarhet genom att paketera kod och dess beroenden till ett *git*[65]-repository. Projektet definieras genom en YAML-fil där beroenden specificeras genom en *Conda*[66]-miljö eller Docker-containrar [55], [56].
- *MLflow Models* baseras på ett generiskt format för att paketera modeller, kod och databeroenden på sådant sätt att kompabilitet med diverse driftsmiljöer uppnås, men även med olika ML-bibliotek [56].
- *MLflow Model Registry* är ett nytt tillskott inriktat på att främja samarbete genom att centralisera lagringen för verksamheten [56].

5.3.2 Comet

Comet[67] är en kommersiell plattform inriktad på datavetenskap och ML. Huvudsakliga funktioner inkluderar att kartlägga, jämföra, förklara och optimera modeller för hela dess livscykel. Därmed inkluderas kedjan från experimenthantering till övervakning i produktion. Lösningarna kan baseras på molntjänster eller lokala servrar [57].

Det huvudsakliga stödet tillhandahålls för Python, men inkluderar även diverse andra programmeringsspråk. Genom ett bibliotek aktiveras loggning av exempelvis hyperparametrar, programkod, och utvärderingsmått genom enstaka rader kod. Plattformen erbjuder sedan stöd för att övervaka och dela resultaten för modeller i realtid genom dynamiska visualiseringar och rapporter. Centralt för plattformen är att tillhandahålla anpassningsbarhet för användaren så att egna bibliotek kan integreras med Comets tjänster för modellhantering [57].

---

[64] https://mlflow.org/ (besökt december 2022)

[65] https://git-scm.com/ (besökt december 2022)

[66] https://conda.io/ (besökt december 2022)

[67] https://www.comet.com/ (besökt december 2022)

5.3.3 DataRobot

DataRobot[68] är en AI-systemplattform som jämförelsevis snarare bör beskrivas som en större uppsättning verktyg som svarar upp mot en tydlig ambition att tillhandahålla en end-to-end lösning. Genom att erbjuda flexibilitet och kontroll sägs processen att gå från idé till produktion accelereras för AI-lösningar. Plattformen förenklar samarbete genom att samla projektresurserna på ett ställe, varför tredjeparts lösningar är vanligt förekommande, vilka sammanfogas genom plattformens GUI [52], [58].

Plattformens verktyg kan främst delas upp i databearbetning, experimentering och produktionssättning. Vidare har plattformen olika delar för att tillhandahålla stöd för användare med varierande kompetenser och behov. Det finns stöd för diverse programmeringsspråk och digitala anteckningsblock, och plattformen hanterar infrastrukturen för att bespara användaren tid [52], [58].

DataRobot verkar för att demokratisera tillgången till AI. Som en del i det tillhandahålls en AutoML-tjänst som åt användaren hanterar variabelframställning, modellselektering och finjustering. Det finns även verktyg inriktade på förklaringsbar AI (XAI, se avsnitt 6.1) och för att kartlägga bias utifrån ett etiskt perspektiv [52], [58].

[68] https://www.datarobot.com/ (besökt december 2022)

6 Övrigt

Utöver redogjorda grupperingar av området inbegriper utvärdering av AI-metoder andra aspekter som bör betraktas. I avsnitt 6.1 noteras vikten av förklaringsbarhet utöver prestanda. I avsnitt 6.2 noteras vikten och svårigheten att välja rätt definition av prestanda. I avsnitt 6.3 introduceras datamätningar som koncept. I avsnitt 6.4 exemplifieras några fall med användning av tävlingar som ett verktyg för utvärdering och kartläggning. I avsnitt 6.5 redogörs för ytterligare programmeringsbibliotek som belyser diverse andra nyttiga aspekter för utvärdering. I avsnitt 6.6 redogörs en kartläggning av det uppgiftsorienterade området avvikelseupptäckt med tillhörande material för utvärdering.

6.1 Förklarbar AI

Darpa[69] lanserade 2017 projektet *eXplainable AI* (XAI) som en konsekvens av lovande framsteg inom ML men där prestandaökningen genererar mer komplexa modeller vilka beskrivs som icke-transparanta. Ett vanligt tillkortakommande för systemen är oförmågan att förklara slutsatser, beslut och handlingar för den mänskliga användaren. XAI utgör en grund i arbetet att förstå och införliva förtroende för AI-system genom särskilt utvecklad metodik [59]. Aspekten att förstå ett system utgör sannolikt i många tillämpningar en viktig del i utvärdering och jämförelse av metoder och modeller.

Ett AI-systems förklarbarhet handlar både om tekniska och mänskliga faktorer och kan mätas exempelvis genom 1) förklaringarnas utformning (exempelvis hur komplett eller detaljerad), 2) användares tillfredställelse med förklaringarna (som fångar användarnas känsla av tillfredsställelse), och 3) hur väl förklaringarna stimulerar användarens nyfikenhet [60].

6.2 Utvärderingsmått

Utvärderingsmått (exempelvis träffsäkerhet) kan vara ett sätt att kvantifiera utvecklingen inom ett fält. Men de kan även användas för att avgöra vilken modell som är mest lämplig avseende en specifik tillämpnings kravställning. Allmänt erkända utvärderingsmått som används för att utvärdera *state of the art* behöver inte vara mest lämpade för en specifik tillämpning. Som stöd för att välja rätt utvärderingsmått lanseras tjänster för att samla diverse mått och verifiera implementationen, exempelvis Python-biblioteket *evaluate*[70] [16].

[69] https://www.darpa.mil/ (besökt december 2022)

[70] https://github.com/huggingface/evaluate (besökt januari 2023)

6.3 Datamätningar

Datamätningar består av kvantifieringar av data för att förstå dess komposition och egenskaper. Nyttan är mångsidig, däribland skapande av nya datamängder, dokumentering av befintliga datamängder, och ökad förståelse av system. Därav har datamätningar varit centrala för ML under lång tid, men detta till trots utgörs området av begränsad konsensus gällande vad som kan och ska mätas, och hur mätningen ska utföras. På senare tid har diverse verktyg lanserats för att mäta kvaliteten på datamängder, men omfattningen är begränsad vad gäller kvantifieringsmått och applicerbara modaliteter. Föreslagna mått kan kategoriseras som [61]:

- *Distans*: mäter separation inom datamängden, exempelvis Euklidiska avståndet[71].
- *Densitet*: mäter kompakthet för datamängden eller delmängder, exempelvis k-närmaste grannar[72].
- *Mångfald*: mäter hur homogena datapunkter är givet en datamängd, exempelvis entropi[73].
- *Tendens*: vilket inbegriper mått relaterade till datafördelningen, exempelvis skevhet[74].
- *Association*: mäter relationen mellan objekt i en datamängd, exempelvis Pearson korrelation[75].

Respektive kategori består även av modalitetsspecifika mått som kvantifierar egenskaper för exempelvis bilder eller text. Andra mått som faller utanför taxonomin är exempelvis mått för redundans och brus [61].

6.4 Tävlingar

Tävlingar är särskilt inriktade på att lösa en uppgift. Men tävlingsmomentet ställer krav på rättvisande utvärderingar och jämförelser. Konferensen NeurIPS, som introducerades i avsnitt 3.2, har en särskild gren som utgörs av tävlingar. För sjätte året i ordningen arrangerades 2022 års upplaga vilken bestod av 25 tävlingar[76]. Området är värdefullt eftersom det ger ett gemensamt fokus på några särskilda problem (ofta inom maskininlärning), samt ger prestandamätningar av

[71] https://en.wikipedia.org/wiki/Euclidean_distance (besökt mars 2023)

[72] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm (besökt mars 2023)

[73] https://en.wikipedia.org/wiki/Entropy_(information_theory) (besökt mars 2023)

[74] https://en.wikipedia.org/wiki/Skewness (besökt mars 2023)

[75] https://en.wikipedia.org/wiki/Pearson_correlation_coefficient (besökt mars 2023)

[76] https://neurips.cc/Conferences/2022/CompetitionTrack (besökt december 2022)

metoder och tillgängliggör datamängder eller interaktiva miljöer [62]. Flera av tävlingarna tillämpar CodaLab (se avsnitt 3.3.3) för resultatlistor.

Angry birds AI competition[77] är en annan tävling som hållits årligen sedan 2012 och som fokuserar på intelligenta agenter vilka spelar datorspelet *Angry birds*[78]. Det långsiktiga målet är att utveckla agenter som spelar bättre än de bästa mänskliga spelarna. Utmaningen för agenterna består i att lära sig förstå vilka konsekvenser handlingar får i spelets (simulerade) fysiska miljö, samtidigt som det finns ett obegränsat antal handlingsalternativ att välja mellan. En skicklig agent i det här sammanhanget förväntas ha skaffat sig förmågor som är användbara även i vår riktiga fysiska värld.

En veteran i tävlingssammanhang är *RoboCup*[79] som startade 1997 och som omfattar olika kategorier med fysiska agenter (robotar) respektive simulerade agenter. Sammanhanget här var från början fotbollsspel och förutom den enskilda agentens utmaningar finns också behovet av kommunikation och samordning mellan agenterna i fotbollslagen. På senare år har tävlingarna utökats med en allvarligare kategori som handlar om robotar för räddningsuppdrag.

6.5 Programmeringsbibliotek

Adversarial Robustness Toolbox[80] (ART) är ett Python-bibliotek inriktat på ML-säkerhet. Verktyget kan exempelvis användas för utvärdering och försvar av modeller mot attacker för uthämtning av information eller vilseledning. ART kan användas för att mäta robustheten för ML-modeller [63]. Säkerhetsaspekten är troligen en komponent som bör betraktas vid helomfattande utvärderingar av AI.

Open Neural Network Exchange[81] (ONNX) är ett verktyg som strävar efter att möjliggöra samarbete mellan olika verktyg för DL genom att tillhandahålla definitioner av grafiska beräkningsmodeller, operatorer och standarddatatyper. Några ramverk som ingår är TensorFlow, PyTorch och Matlab. ONNX är ett samarbete mellan aktörer och verktyget baseras på öppen källkod [64]. Verktyget bedöms vara intressant vid eventuell utveckling av en utvärderingsplattform med hänvisning till möjligheten att sammanlänka olika programmeringsspråk och bibliotek.

CheckList[82] är ett Python-bibliotek för att låta en användare snabbare generera utmanande testfall för språkteknologimodeller för att tillåta utvärdering gällande

[77] http://aibirds.org/ (besökt februari 2023)

[78] Utvecklat av det finska företaget Rovio (https://www.angrybirds.com/, besökt februari 2023)

[79] https://www.robocup.org/ (besökt februari 2023)

[80] https://adversarial-robustness-toolbox.org/ (besökt december 2022)

[81] https://onnx.ai/ (besökt december 2022)

[82] https://github.com/marcotcr/checklist (besökt december 2022)

generaliserbarhet. Exempel på generering av testfall kan vara tillämpning av synonymer eller omformuleringar av meningar för att undersöka förmågan att hantera negationer eller mer avancerade syftningar [65].

rliable[83] är ett Python-bibliotek inriktat på RL för att tillhandahålla metodik för att styrka konfidensen gällande jämförelser av resultat genom enbart en handfull repetitioner. Nyttan består i att jämförelse av resultat genom punktestimeringar (till exempel träffsäkerhet) bortser från osäkerhetsaspekten. Samtidigt är det ofta en väldigt tidskrävande process att upprepa RL experiment. Centralt för biblioteket är samplingsmetoder för att kvantifiera resultaten och inspektera, ofta visuellt, osäkerhetsaspekten [66].

6.6 Avvikelseupptäckt

Anomaly detection (sv. *avvikelseupptäckt)* är samlingsbegreppet för metoder inriktade på att identifiera mönster i data som avviker från det förväntade beteendet vilket beskrivs som det normala. Avvikelserna benämns vanligen med begrepp så som anomalier, utliggare eller utstickare. Tillämpningsområden för avvikelseupptäckt är många, däribland militärövervakning av fiendens aktiviteter [67].

Avsnitt 6.6.1 beskriver metodik för avvikelseupptäckt, följt av avsnitt 6.6.2 som beskriver diverse resurser för tillhörande utvärdering.

6.6.1 Metodik för avvikelseupptäckt

För avvikelseupptäckt kan fundamentalt olika angreppssätt tillämpas, där oövervakad inlärning är vanligast, kontrasterat av övervakad och semi-övervakad inlärning. En anledning tillskrivs kostnaden för manuell annotering av data. Men lärande baserat på annotering är även begränsande eftersom det avvikande beteendet ofta är okänt eller föränderligt varpå nya typer av avvikelser framträder vilka inte nödvändigtvis inbegrips av den annoterade datamängden [67].

Ett viktigt antagande för oövervakad inlärning för avvikelseupptäckt är att de normala observationerna är i betydande majoritet till avvikelserna i testdatamängden. Anomalier kan differentieras i tre kategorier [67]:

- *Punktanomalier* vars värden avviker till den betraktade datamängdens gränser.
- *Kontextanomalier* vars värden avviker i förhållande till ett tidsberoende till exempel.

[83] https://github.com/google-research/rliable (besökt december 2022)

- G*ruppanomalier* vars individuella observationer inte nödvändigtvis är avvikande, men där samlingen observationer är avvikande i relation till helheten.

Förekomst av föreslagna metoder i litteraturen inbegriper exempelvis:

- *Associationsanalys* vilket genererar regler för kategorisk data [67].
- *Klustring* vilket grupperar datapunkter med likheter enligt något kriterium [67].
- *Spektrala tekniker* vilket baseras på projektion till lägre dimension där egenskaper av intresse framträder [67].
- *Djupinlärning* vilket lämpar sig särskilt för hantering av bilder och sekventiell data [68].

6.6.2 Utvärdering för avvikelseupptäckt

Vad som är att betrakta som en avvikelse varierar mellan olika uppgifter. Däremot är obalanserade datamängder en gemensam nämnare, där förekomsten av avvikelser är i minoritet jämfört med det normala. Ett sådant förhållande kräver att ML-modeller utvärderas och jämförs med anpassad metodik. De tillämpade utvärderingsmåtten bör exempelvis inriktas på förmågan gällande minoritetsklassen snarare än modellens övergripande förmåga. Relaterat är värderingen av kostnaden för falsklarm (typ I-fel) respektive missar (typ II-fel). Föreslagna utvärderingsmått är exempelvis *precision* och *recall*, eller det harmoniska medelvärdet av dessa mått benämnt *F1-score* [69].

Utvärderingsmått är ett exempel för det som beskrivs som intern validitetsproblematik, medan den externa valideringsproblematiken relaterar förmågan att generalisera resultaten till andra förutsättningar och domäner. Aspekter att betrakta är exempelvis den använda datamängdens representativitet för uppgiften att lösa i praktiken [69].

Det finns resurser vilka tillhandahåller publika datamängder för riktmärkning av avvikelseupptäckt. Eftersom uppgiften inte är metodorienterad utan uppgiftsorienterad så är täckningsgraden för de olika lösningarna varierande, vilket även återspeglas i förekomsten av olika modaliteter av data. Några exempel på tillgängliga resurser är:

- *Numenta Anomaly Benchmark*[84] som består av observerade tidsserier med avvikelser annoterade [70].
- *ADBench*[85] som består delvis av en sammanställning av diverse befintliga publika datamängder, observerade som syntetiska, med avvikelser

[84] https://github.com/numenta/NAB (besökt januari 2023)

[85] https://github.com/Minqi824/ADBench (besökt januari 2023)

annoterade. Huvudsakligen består resurserna av tabellformsdata men det förekommer inslag för datorseende och språkteknologi [71].

- *SegmentMeIfYouCan*[86] som består av segmenterade datamängder, annoterade avseende avvikande objekt i kontexten av bilvägar. Resursen består även av ledarlistor med länkar till studier som använt de tillhandahållna datamängderna [72].

[86] https://segmentmeifyoucan.com/ (besökt januari 2023)

7 Rekommendationer

Den interaktiva aspekten angående riktmärkning för de jämförelseplattformar som studeras i projektet bedöms som nyttig. Det gäller dels möjligheten att kunna filtrera och sortera, men framförallt att dynamiskt kunna sammanväga utvärderingsmått enligt användarens premisser. Det identifieras även en nytta att med enkelhet kunna skapa nya egna scenarier vilka kan användas för att verifiera en tränad modell. Nyttan består exempelvis av att angripa problematiken med god uppvisad prestanda på riktmärkesuppgifter men misslyckande på enkla verkliga uppgifter. Den omgivningsbaserade utvärderingen är främst inriktad på förstärkningsinlärning (RL), men som redovisats inbegrips fler domäner. På sikt bedöms den här typen av resurser vara mest intressant då ambitionsnivån berör AGI, men i dagsläget krävs kombinationer av riktmärkesverktyg för utvärdering inom fältet som helhet.

Resultaten av framförallt maskininlärning (ML) beror till stor del på kvaliteten av data, och huruvida data inkorporerat i en modell är representativ för den uppgift som ämnas lösas. Med en utgångspunkt från en relativt generisk datamängd som grund blir således utvärderingen av en sådan modell generisk, och därmed främst lämplig för utvärdering av utveckling och trender inom fältet i allmänhet i relation till datamängden samt dess egenskaper beskrivna av utvärderingsmetoden. Men i ytterst syfte att utvärdera landvinningarnas lämplighet för den aktuella domänen, snarare än i allmänhet, väcks ett behov av egenutvecklade datamängder som kan utgöra en del i en intern riktmärkning.

Omfattningen av MLOps är betydligt större än det här arbetets fokus, men bistår dock med erfarenheter och god tillämpning av arbetssätt. Möjligen kan fördjupade studier även identifiera ett redan befintligt verktyg som passande. Från den genomförda kartläggningen identifieras följande egenskaper som särskilt viktiga för utvärdering av AI-system:

- Versionshantering av data och interaktiva miljöer.
- Strömlinjeformad dokumentering gällande framställning och användning av modeller.
- Katalogisering av experiment för reproducerbarhet och jämförelser. Det möjliggör exempelvis utvärdering genom visualiseringar.
- Integrering av flera användare i samma system vilket utöver uppgiftsorienterat samarbete även möjliggör förenklad delning av utvecklade generella utvärderingsmetoder.
- Flexibilitet avseende exempelvis infrastruktur och programmeringsbibliotek för att möta områdets imponerande utvecklingstakt.

Avslutningsvis konkluderar arbetet att området om hur man ska jämföra och utvärdera AI-system är ett stort omfattande område som i dagsläget kräver domän-

specifika resurser då ett universellt verktyg saknas. Emellertid finns det beröringspunkter där standarder för exempelvis dokumentering och versionshantering kan vara av nytta.

Referenslista

[1] ISO, ”ISO/IEC TR 29119-11:2020(en), Software and systems engineering — Software testing — Part 11: Guidelines on the testing of AI-based systems”, *International Organization for Standardization*. https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:29119:-11:ed-1:v1:en (åtkomstdatum 20 oktober 2022).

[2] A. Avizienis, J.-C. Laprie, B. Randell, och C. Landwehr, ”Basic concepts and taxonomy of dependable and secure computing”, *IEEE Trans. Dependable Secure Comput.*, vol. 1, nr 1, s. 11–33, jan. 2004, doi: 10.1109/TDSC.2004.2.

[3] S. Arora och B. Barak, *Computational Complexity: A Modern Approach*. Cambridge University Press, 2009. doi: 10.1017/CBO9780511804090.

[4] S. Zilberstein, ”Using Anytime Algorithms in Intelligent Systems”, *AI Mag.*, vol. 17, nr 3, Art. nr 3, mar. 1996, doi: 10.1609/aimag.v17i3.1232.

[5] A. M. Turing, ”I.—Computing Machinery and Intelligence”, *Mind*, vol. LIX, nr 236, s. 433–460, okt. 1950, doi: 10.1093/mind/LIX.236.433.

[6] B. Goertzel, ”Artificial General Intelligence: Concept, State of the Art, and Future Prospects”, *J. Artif. Gen. Intell.*, vol. 5, nr 1, s. 1–48, dec. 2014, doi: 10.2478/jagi-2014-0001.

[7] ”Testbed”, *Wikipedia*. 09 maj 2022. Åtkomstdatum: 20 oktober 2022. [Online]. Tillgänglig vid: https://en.wikipedia.org/w/index.php?title=Testbed&oldid=1087005729

[8] K. Kyprianidis och E. Dahlquist, *AI and Learning Systems - Industrial Applications and Future Directions*. 2021. doi: 10.5772/intechopen.85833.

[9] R. Merritt, ”What is MLOps?”, *NVIDIA Blog*, 03 september 2020. https://blogs.nvidia.com/blog/2020/09/03/what-is-mlops/ (åtkomstdatum 15 november 2022).

[10] ”Benchmarking”, *Wikipedia*. 12 september 2022. Åtkomstdatum: 20 oktober 2022. [Online]. Tillgänglig vid: https://en.wikipedia.org/w/index.php?title=Benchmarking&oldid=1109899874

[11] D. Zhang *m.fl.*, ”The AI Index 2021 Annual Report”. arXiv, 08 mars 2021. doi: 10.48550/arXiv.2103.06312.

[12] B. Koch, E. Denton, A. Hanna, och J. G. Foster, ”Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research”, s. 13.

[13] D. Yadav, R. Jain, H. Agrawal, och P. Chattopadhyay, ”EvalAI: Towards Better Evaluation of AI Agents”, s. 3.

[14] N. Polyzotis och M. Zaharia, ”What can Data-Centric AI Learn from Data and ML Engineering?” arXiv, 13 december 2021. Åtkomstdatum: 01 november 2022. [Online]. Tillgänglig vid: http://arxiv.org/abs/2112.06439

[15] ”Reproducible Research in Computational Science”. https://www.science.org/doi/10.1126/science.1213847 (åtkomstdatum 06 december 2022).

[16] L. von Werra *m.fl.*, ”Evaluate & Evaluation on the Hub: Better Best Practices for Data and Model Measurements”. arXiv, 06 oktober 2022. Åtkomstdatum: 26 oktober 2022. [Online]. Tillgänglig vid: http://arxiv.org/abs/2210.01970

[17] C. de la Torre, B. Wagner, och M. Rousos, *.NET Microservices: Architecture for Containerized .NET Applications*. Microsoft Developer Division, .NET and Visual Studio product teams, 2022. Åtkomstdatum: 01 november 2022. [Online]. Tillgänglig vid: https://dotnet.microsoft.com/en-us/download/e-book/microservices-architecture/pdf

[18] T. Gebru *m.fl.*, ”Datasheets for datasets”, *Commun. ACM*, vol. 64, nr 12, s. 86–92, dec. 2021, doi: 10.1145/3458723.

[19] B. Hutchinson *m.fl.*, ”Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure”. arXiv, 29 januari 2021. Åtkomstdatum: 10 november 2022. [Online]. Tillgänglig vid: http://arxiv.org/abs/2010.13561

[20] M. Pushkarna, A. Zaldivar, och O. Kjartansson, ”Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI”, i *2022 ACM Conference on Fairness, Accountability, and Transparency*, Seoul Republic of Korea, juni 2022, s. 1776–1826. doi: 10.1145/3531146.3533231.

[21] M. Mitchell *m.fl.*, ”Model Cards for Model Reporting”, i *Proceedings of the Conference on Fairness, Accountability, and Transparency*, jan. 2019, s. 220–229. doi: 10.1145/3287560.3287596.

[22] OECD, *AI and the Future of Skills, Volume 1: Capabilities and Assessments*. OECD, 2021. doi: 10.1787/5ee71f34-en.

[23] M. Mazumder *m.fl.*, ”DataPerf: Benchmarks for Data-Centric AI Development”. arXiv, 20 juli 2022. Åtkomstdatum: 24 oktober 2022. [Online]. Tillgänglig vid: http://arxiv.org/abs/2207.10062

[24] C. G. Northcutt, A. Athalye, och J. Mueller, ”Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks”. arXiv, 07 november 2021. Åtkomstdatum: 24 oktober 2022. [Online]. Tillgänglig vid: http://arxiv.org/abs/2103.14749

[25] L. Aroyo, M. Lease, P. Paritosh, och M. Schaekermann, ”Data excellence for AI: why should you care?”, *Interactions*, vol. 29, nr 2, s. 66–69, mar. 2022, doi: 10.1145/3517337.

[26] S. Yeung och J. Vanschoren, ”Announcing the NeurIPS 2021 Datasets and Benchmarks Track”, *Medium*, 08 april 2021. https://neuripsconf.medium.com/announcing-the-neurips-2021-datasets-and-benchmarks-track-644e27c1e66c (åtkomstdatum 09 november 2022).

[27] P. Mattson *m.fl.*, ”MLPerf: An Industry Standard Benchmark Suite for Machine Learning Performance”, *IEEE Micro*, vol. 40, nr 2, s. 8–16, mar. 2020, doi: 10.1109/MM.2020.2974843.

[28] D. Kiela *m.fl.*, ”Dynabench: Rethinking Benchmarking in NLP”, i *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online, 2021, s. 4110–4124. doi: 10.18653/v1/2021.naacl-main.324.
[29] Z. Epstein *m.fl.*, ”TuringBox: An Experimental Platform for the Evaluation of AI Systems”, i *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI’18)*, Stockholm, Sweden, juli 2018, s. 5826–5828. doi: 10.24963/ijcai.2018/851.
[30] K. Ethayarajh och D. Jurafsky, ”Utility is in the Eye of the User: A Critique of NLP Leaderboards”. arXiv, 03 mars 2021. Åtkomstdatum: 21 oktober 2022. [Online]. Tillgänglig vid: http://arxiv.org/abs/2009.13888
[31] Z. Ma *m.fl.*, ”Dynaboard: An Evaluation-As-A-Service Platform for Holistic Next-Generation Benchmarking”, s. 19.
[32] T. Thrush *m.fl.*, ”Dynatask: A Framework for Creating Dynamic AI Benchmark Tasks”. arXiv, 04 april 2022. Åtkomstdatum: 21 oktober 2022. [Online]. Tillgänglig vid: http://arxiv.org/abs/2204.01906
[33] A. Pavao *m.fl.*, ”CodaLab Competitions: An open source platform to organize scientific challenges”, s. 6.
[34] Z. Xu *m.fl.*, ”Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform”, *Patterns*, vol. 3, nr 7, s. 100543, juli 2022, doi: 10.1016/j.patter.2022.100543.
[35] A. Srivastava *m.fl.*, ”Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models”. arXiv, 10 juni 2022. doi: 10.48550/arXiv.2206.04615.
[36] A. Wang *m.fl.*, ”SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems”, i *Advances in Neural Information Processing Systems*, 2019, vol. 32. Åtkomstdatum: 17 november 2022. [Online]. Tillgänglig vid: https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html
[37] M. Mazumder *m.fl.*, ”Multilingual Spoken Words Corpus”, *Proc. Neural Inf. Process. Syst. Track Datasets Benchmarks*, vol. 1, dec. 2021, Åtkomstdatum: 17 november 2022. [Online]. Tillgänglig vid: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/fe131d7f5a6b38b23cc967316c13dae2-Abstract-round2.html
[38] J. F. Gemmeke *m.fl.*, ”Audio Set: An ontology and human-labeled dataset for audio events”, i *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.
[39] P. Dendorfer *m.fl.*, ”MOTChallenge: A Benchmark for Single-Camera Multiple Target Tracking”, *Int. J. Comput. Vis.*, vol. 129, nr 4, s. 845–881, apr. 2021, doi: 10.1007/s11263-020-01393-0.
[40] V. Pătrăucean *m.fl.*, ”Perception Test: A Diagnostic Benchmark for Multimodal Models”, s. 22.
[41] A. H. Miller *m.fl.*, ”ParlAI: A Dialog Research Software Platform”. arXiv, 08 mars 2018. Åtkomstdatum: 17 november 2022. [Online]. Tillgänglig vid: http://arxiv.org/abs/1705.06476

[42] G. Brockman *m.fl.*, ”OpenAI Gym”. arXiv, 05 juni 2016. Åtkomstdatum: 15 november 2022. [Online]. Tillgänglig vid: http://arxiv.org/abs/1606.01540
[43] The Farama Foundation, ”Announcing The Farama Foundation - The future of open source reinforcement learning”, *The Farama Foundation*, 25 oktober 2022. https://farama.org/Announcing-The-Farama-Foundation (åtkomstdatum 16 november 2022).
[44] I. Osband *m.fl.*, ”Behaviour Suite for Reinforcement Learning”. arXiv, 14 februari 2020. Åtkomstdatum: 17 november 2022. [Online]. Tillgänglig vid: http://arxiv.org/abs/1908.03568
[45] M. Crosby, B. Beyret, M. Shanahan, J. Hernández-Orallo, L. Cheke, och M. Halina, ”The Animal-AI Testbed and Competition”, i *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, aug. 2020, s. 164–176. Åtkomstdatum: 21 november 2022. [Online]. Tillgänglig vid: https://proceedings.mlr.press/v123/crosby20a.html
[46] M. Johnson, K. Hofmann, T. Hutton, och D. Bignell, ”The Malmo Platform for Artificial Intelligence Experimentation”, s. 2.
[47] L. M. Eberding, K. R. Thórisson, A. Sheikhlar, och S. P. Andrason, ”SAGE: Task-Environment Platform for Evaluating a Broad Range of AI Learners”, i *Artificial General Intelligence*, vol. 12177, B. Goertzel, A. I. Panov, A. Potapov, och R. Yampolskiy, Red. Cham: Springer International Publishing, 2020, s. 72–82. doi: 10.1007/978-3-030-52152-3_8.
[48] K. R. Thorisson, J. Bieger, S. Schiffel, och D. Garrett, ”Towards Flexible Task Environments for Comprehensive Evaluation of Artificial Intelligent Systems & Automatic Learners”, s. 10.
[49] A. Dorri, S. S. Kanhere, och R. Jurdak, ”Multi-Agent Systems: A Survey”, *IEEE Access*, vol. 6, s. 28573–28593, 2018, doi: 10.1109/ACCESS.2018.2831228.
[50] C. Xue, V. Pinto, C. Gamage, E. Nikonova, P. Zhang, och J. Renz, ”Phy-Q as a measure for physical reasoning intelligence”, *Nat. Mach. Intell.*, vol. 5, nr 1, Art. nr 1, jan. 2023, doi: 10.1038/s42256-022-00583-4.
[51] D. Jeffries, ”Why We Started the AIIA and What It Means for the Rapid Evolution of the Canonical Stack of Machine Learning”, *AI Infrastructure Alliance*, 07 januari 2022. https://ai-infrastructure.org/why-we-started-the-aiia-and-what-it-means-for-the-rapid-evolution-of-the-canonical-stack-of-machine-learning/ (åtkomstdatum 14 november 2022).
[52] AI Infrastructure Alliance, ”AI Infrastructure Ecosystem 2022”, 20220721001. Åtkomstdatum: 15 november 2022. [Online]. Tillgänglig vid: https://ai-infrastructure.org/ai-infrastructure-ecosystem-report-of-2022/
[53] AI Infrastructure Alliance, ”AI Infrastructure Landscape”, *AI Infrastructure Alliance*. https://ai-infrastructure.org/ai-infrastructure-landscape/ (åtkomstdatum 14 november 2022).

[54] M. Gualtieri och R. Curran, ”The Forrester Wave™: AI/ML Platforms, Q3 2022”, *Forrester*. https://reprints2.forrester.com/#/assets/2/1593/RES176365/report (åtkomstdatum 01 november 2022).
[55] A. Chen *m.fl.*, ”Developments in MLflow: A System to Accelerate the Machine Learning Lifecycle”, i *Proceedings of the Fourth International Workshop on Data Management for End-to-End Machine Learning*, Portland OR USA, juni 2020, s. 1–4. doi: 10.1145/3399579.3399867.
[56] M. Zaharia *m.fl.*, ”Accelerating the Machine Learning Lifecycle with MLflow”, s. 7.
[57] Comet, ”Comet Docs”. https://www.comet.com/docs/v2/ (åtkomstdatum 24 november 2022).
[58] DataRobot, ”For Data Scientists, By Data Scientists”, *DataRobot AI Cloud*. https://www.datarobot.com/solutions/data-scientists/ (åtkomstdatum 14 november 2022).
[59] D. Gunning och D. Aha, ”DARPA’s Explainable Artificial Intelligence (XAI) Program”, *AI Mag.*, vol. 40, nr 2, s. 44–58, juni 2019, doi: 10.1609/aimag.v40i2.2850.
[60] R. Hoffman, S. T. Mueller, G. Klein, och J. Litman, ”Metrics for Explainable AI: Challenges and Prospects”, *ArXiv*, dec. 2018, Åtkomstdatum: 09 februari 2023. [Online]. Tillgänglig vid: https://www.semanticscholar.org/paper/Metrics-for-Explainable-AI%3A-Challenges-and-Hoffman-Mueller/be711f681580d3a02c8bc4c4dab0c7a043f4e1d2
[61] M. Mitchell *m.fl.*, ”Measuring Data”. arXiv, 09 december 2022. Åtkomstdatum: 16 december 2022. [Online]. Tillgänglig vid: http://arxiv.org/abs/2212.05129
[62] S. Ghalebikesabi, ”Announcing NeurIPS 2022 Competitions”, *NeurIPS Blog*. https://blog.neurips.cc/2022/05/31/neurips-2022-competitions-announced/ (åtkomstdatum 09 november 2022).
[63] Adversarial Robustness Toolbox, ”User Guide”. https://adversarial-robustness-toolbox.readthedocs.io/en/latest/ (åtkomstdatum 09 december 2022).
[64] ONNX, ”About”. https://onnx.ai/about.html (åtkomstdatum 09 december 2022).
[65] M. T. Ribeiro, T. Wu, C. Guestrin, och S. Singh, ”Beyond Accuracy: Behavioral Testing of NLP Models with CheckList”, i *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, juli 2020, s. 4902–4912. doi: 10.18653/v1/2020.acl-main.442.
[66] R. Agarwal, M. Schwarzer, P. S. Castro, A. C. Courville, och M. Bellemare, ”Deep Reinforcement Learning at the Edge of the Statistical Precipice”, i *Advances in Neural Information Processing Systems*, 2021, vol. 34, s. 29304–29320. Åtkomstdatum: 17 november 2022. [Online]. Tillgänglig vid: https://proceedings.neurips.cc/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html

[67] V. Chandola, A. Banerjee, och V. Kumar, ”Anomaly Detection: A Survey”, *ACM Comput Surv*, vol. 41, juli 2009, doi: 10.1145/1541880.1541882.

[68] R. Chalapathy och S. Chawla, ”Deep Learning for Anomaly Detection: A Survey”. arXiv, 23 januari 2019. Åtkomstdatum: 23 november 2022. [Online]. Tillgänglig vid: http://arxiv.org/abs/1901.03407

[69] R. Guizzardi, J. Ralyté, och X. Franch, Red., *Research Challenges in Information Science: 16th International Conference, RCIS 2022, Barcelona, Spain, May 17–20, 2022, Proceedings*, vol. 446. Cham: Springer International Publishing, 2022. doi: 10.1007/978-3-031-05760-1.

[70] S. Ahmad, A. Lavin, S. Purdy, och Z. Agha, ”Unsupervised real-time anomaly detection for streaming data”, *Neurocomputing*, vol. 262, s. 134–147, nov. 2017, doi: 10.1016/j.neucom.2017.04.070.

[71] S. Han, X. Hu, H. Huang, M. Jiang, och Y. Zhao, ”ADBench: Anomaly Detection Benchmark”, presenterad vid Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, okt. 2022. Åtkomstdatum: 30 december 2022. [Online]. Tillgänglig vid: https://openreview.net/forum?id=foA_SFQ9zo0

[72] R. Chan *m.fl.*, ”SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation”. arXiv, 09 november 2021. doi: 10.48550/arXiv.2104.14812.

Bilaga A Begrepp och förkortningar

Förkortning	Svenska	Engelska
AGI	Artificiell generell intelligens	Artificial General Intelligence
AI	Artificiell intelligens	Artificial Intelligence
AIIA	-	AI Infrastructure Alliance
API	Programmeringsgränssnitt	Application Programming Interface
AutoML	Automatisk maskininlärning	Automated Machine Learning
DL	Djupinlärning	Deep Learning
GUI	Grafiskt användargränssnitt	Graphical user interface
ML	Maskininlärning	Machine Learning
NeurIPS	-	(Conference on) Neural Information Processing Systems
RL	Förstärkningsinlärning	Reinforcement learning
UI	Användargränssnitt	User interface
XAI	Förklaringsbar AI	Explainable AI

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