KUNGL. VETENSKAPS-AKADEMIEN
THE ROYAL SWEDISH ACADEMY OF SCIENCES

VETENSKAPEN SÄGER

VETENSKAPEN SÄGER ★ NR 6 ★ APRIL 2024

–om AI

Artificiell intelligens diskuteras intensivt nu. Den tekniska utvecklingen går oerhört fort och för varje dag blir AI en del av livet för allt fler människor. För att fatta kloka beslut kring den här utvecklingen behöver vi förstå vad AI egentligen är, och hur vi kan använda tekniken på bra sätt.

FOTON: ADOBE STOCK OCH COMUZI/© BBC/BETTER IMAGES OF AI/LIKES/CC-BY 4.0

En revolutionerande teknik

Det pågår en revolution just nu. En AI-revolution. Även om grunden för tekniken lades för mer än ett halvt sekel sedan, är det först nu som kunskaperna och datorkraften är tillräckliga för att artificiell intelligens ska bli tillgängligt för många. AI är på väg att bli en del av vardagen. Men hur fungerar AI egentligen, och hur kan den utvecklas i framtiden?

AI, artificiell intelligens, är ett brett begrepp som omfattar många olika tekniker, metoder och användningsområden. I grunden handlar det om att ge maskiner ett slags ”tankeförmågor” som liknar människans, och på så vis göra det möjligt för dem att utföra uppgifter som tidigare bara människor har klarat.

En utmaning när man vill definiera artificiell intelligens är att det inte finns någon allmänt accepterad definition av *mänsklig* intelligens. Datorer kan redan göra många saker som skulle kräva stor mänsklig intelligens, som att lösa svåra matematiska problem. Samtidigt är de ofta dåliga på uppgifter som vi människor klarar av, som att lära oss ett nytt begrepp från abstrakta beskrivningar, eller att med relativt lite träning kategorisera saker. Små barn behöver bara se ett fåtal katter för att förstå att de alla tillhör samma kategori även om katterna har olika färg, storlek och form – en verklig, en ritad, en på foto. Dagens AI däremot behöver träna på väldigt många bilder för att lära sig känna igen en katt.

Det kan vara mer korrekt att säga att AI är system som *beter sig* intelligent, och specificera vad de kan. De kanske kan planera, lära in något eller uppfatta vissa signaler. Ett annat sätt att beskriva AI är som system som tar emot data, analyserar den, fattar beslut och agerar utifrån besluten.

AI kan lära sig, men har begränsningar

Många AI-system kan lära sig och förbättras över tid. Så fungerar till exempel neurala nätverk (se s. 11), som började utvecklas redan på 1950-talet. Då utgick man från en enkel modell av hur forskare trodde att mänskliga neuroner, nervceller, fungerade. I dag kan AI utföra många komplexa uppgifter; tolka mänsklig skrift, läsa av trafiksituationer och bedöma röntgenbilder. I flera fall överträffar systemen människan och de har potential att hjälpa oss med saker som vi aldrig skulle klara av själva.

Forskning och utveckling på AI-området bedrivs både av företag och av akademiska forskare. Det handlar inte bara om att utveckla systemen utan också om att utforska exakt hur de kan lära, se, planera, resonera, hantera språk och mycket mer. Den här typen av forskning har pågått sedan strax efter andra världskriget, men den har aldrig varit så omfattande som nu. Användningen av AI ökar blixtsnabbt och metoderna används i vård, jordbruk, ekonomi, utbildning, kultur, design, myndighetsarbete... På de kommande uppslagen kan du se flera exempel.

Samtidigt har AI fortfarande begränsningar. AI kan identifiera dolda mönster i enorma mängder data på ett sätt som människans blick aldrig klarar av. Men data ger inte en komplett bild av verkligheten. Någon har fattat ett beslut om var data ska hämtas och vilka data som ska ingå. Det är alltså den data som matas in som avgör hur bra systemets klassificeringar och beslut kan bli.

De flesta AI-metoder är baserade på korrelation, det vill säga samband mellan olika faktorer. De är dåliga på att urskilja kausalitet – orsak och verkan. Dessutom saknar AI i allmänhet den uppfattning om sammanhang som människor har och AI har inte samma förståelse för mänskliga samhällen och kulturer. Dagens AI är inte heller så bra på att anpassa sig till helt nya eller oväntade omständigheter.

AI påverkar och kommer att påverka våra liv mer än de flesta tekniker som människan tidigare har uppfunnit och använt sig av. Det är en omställning lika stor som den industriella revolutionen, men den sker på kortare tid. Tekniken har stor potential att ge oss bättre samhällen och bättre liv i framtiden. Samtidigt medför den etiska, juridiska och ekonomiska utmaningar som vi måste hantera. Även det kan du läsa mer om i den här skriften.

AI-forskning i Sverige

- **AKADEMI.** De flesta svenska universitet bedriver forskning och utbildning om AI.
- **NÄRINGSLIV.** I det svenska näringslivet används bland annat avancerade AI-drivna robotar i tillverkningsprocesser, och AI-system för att designa nya läkemedel.
- Forskningsstiftelsen **KNUT OCH ALICE WALLENBERGS STIFTELSE** gör stora satsningar på AI. Bland dem ingår:
  - **WASP,** *Wallenberg AI, Autonomous Systems and Software Program*, det största forskningsprogrammet någonsin i Sverige. 600 doktorander ska ta examen fram till och med år 2030.
  - **WASP-HS**. Ett forskningsprogram med fokus på humaniora och samhälls-vetenskaplig forskning kring AI.
  - **BERZELIUS**. En superdator till stöd för svensk forskning om och användning av maskininlärning. I drift sedan 2021.
- **AI-SWEDEN.** Nationellt center finansierat av statliga Vinnova, fokuserat på AI-tillämpningar.

I den här skriften används mest svenska termer. Vill du söka mer information kan det dock vara lättare att utgå från de engelska orden:

maskininlärning ⟷ machine learning

datorseende ⟷ computer vision

stora språkmodeller ⟷ large language models, LLM

tidsseriedata ⟷ time series data

representationsinlärning ⟷ representation learning

o/övervakad inlärning ⟷ un/supervised learning

förstärkningsinlärning ⟷ reinforcement learning

kategoriserande AI ⟷ discriminative AI

FOTO: BETTER IMAGES OF AI / DATA IS A MIRROR OF US / CC-BY 4.0

"Data is a Mirror of Us". Collage av konstnärerna Anne Fehres och Luke Conroy i projektet AI4media.

AI överallt omkring oss

AI kan snabbt processa stora mängder data och finna mönster som är svåra eller omöjliga för en människa att se. Det går knappt att hitta ett område där AI inte har, eller kommer att ha, enormt inflytande. Maskininlärning, datorseende, modeller för analys och syntes av språk, och AI-drivna program som ger stöd för beslut, förändrar arbetssätt och hela branscher i grunden.

Inom den svenska sjukvården testas just nu system som kan upptäcka bröstcancer. En AI-modell har fått träna på medicinska bilder av friska bröst och bröst med cancervävnad, och kan identifiera subtila mönster kopplade till cancer. Det finns också andra AI-system som används för att automatiskt analysera medicinska bilder och hjälpa vårdpersonalen med diagnos och tolkningar. På så vis går bedömningarna snabbare och blir mer exakta. AI börjar också bli del av olika funktionsstöd som hörapparater, synhjälpmedel och kommunikationshjälpmedel där systemet kan föreslå ord och meningar – så som det redan fungerar i våra smartphones.

AI kan också få stor betydelse i länder där få människor har tillgång till avancerad vård och många bor långt från en läkare. I Indien har forskare tagit fram ett AI-system som kan användas på landsbygden för snabb diagnos av tuberkulos, och i södra Afrika har ett AI-styrt mikroskop utvecklats som detekterar malariaparasiter i blod på några minuter. Det ska användas av vårdpersonal på landsbygden där laboratorier saknas.

Myndighetsarbete, trafik och miljö

Med AI kan myndigheter snabbt och automatiserat översätta information, vilket är viktigt i områden där människor talar många olika språk. Vid kriser kan AI användas för att förvarna om resursbrist och göra

prognoser för hjälpsändningar. När människor flyr naturkatastrofer eller våld kan AI-system analysera vilka de troligaste flyktvägarna blir och visa var nödhjälp behövs.

Trafikplanering är ett annat område där AI:s förmågor är till stor nytta. Svenska forskare har till exempel utvecklat ett system för att förutse faror i flygtrafiken, och hjälpa flygledare att lotsa planen rätt och undvika problem.

Jordbruk, miljö och energi

Lantbrukare runtom i världen har börjat testa AI-system som beräknar behovet av bevattning, gödning och skadedjursbekämpning. Det finns AI som styr skördesystem i växthus, detekterar läckor i bevattningssystem och driver drönare som övervakar odlingar och betande djur.

Även i naturvård testas AI. I Storbritannien har forskare använt AI-kontrollerade kameror och mikrofoner för att identifiera och inventera vilda djur och fåglar i naturen. Med den informationen får man bättre överblick över vilka arter som är hotade och i vilka miljöer de rör sig.

AI spelar också en stor roll i omställningen till mer hållbara energisystem. AI kan styra så kallade smarta elnät, där tillgången anpassas till efterfrågan i olika områden för att minska risken för strömavbrott. Den kan också göra beräkningar av när energisystem kan väntas behöva underhåll, när och hur energi behöver lagras för att kunna fördelas på bästa sätt, och vilken energiproduktion som förnybara källor kommer att ge givet bland annat väderprognoserna.

AI förändrar forskningen

AI är betydligt bättre än människan på att finna mönster, och har en enorm kapacitet för att snabbt och systematiskt skapa och testa olika varianter av något. Det är en stor tillgång inom många forskningsfält. AI-modeller används bland annat inom kemi och fysik för att simulera fysiska fenomen eller komplexa samspel mellan molekyler, vilket gör det möjligt att både förutse och optimera olika processer. I materialvetenskap används AI för att snabbt testa strukturer och på så vis öka tempot i utvecklingen av nya material.

Inom bioinformatik simulerar AI interaktion mellan olika proteiner, och avslöjar sjukdomsmekanismer. Kraftfulla datorer och AI-system kan analysera enorma datamängder på ett vis som skulle ha varit helt otänkbart för bara några år sedan. På så vis har det blivit möjligt att göra vetenskapliga upptäckter på helt nya sätt.

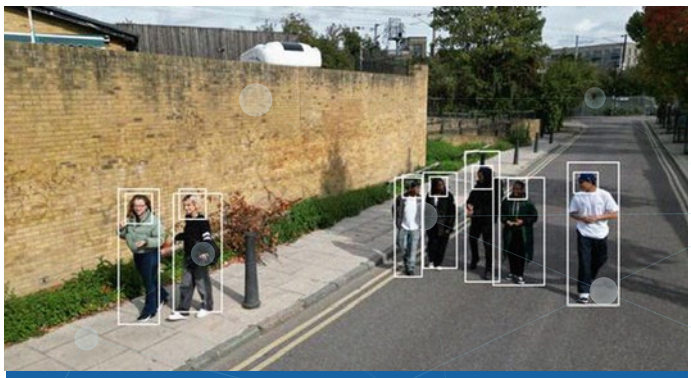


FOTO: COMUZI / © BBC / BETTER IMAGES OF AI / SURVEILLANCE VIEW A / CC-BY 4.0

Bildanalys

När datorseende och AI kombineras blir maskiner allt bättre på att uppfatta och tolka visuell information. Det används bland annat i mobilers och övervakningskamerors ansiktsigenkänning, i självkörande fordon och vid bildanalyser i vården där AI med stor träffsäkerhet kan urskilja cancerceller i vävnadsprover.

Tidsserieprediktion

Ordet *prediktion* inom AI-området innebär att beräkningsmodeller används för att utifrån tidigare data ”gissa” – prognostisera – framtida data. AI-system analyserar stora datamängder för att identifiera trender och relationer, och kan på så vis förutspå kommande händelser och värden. Det här används inom bland annat finansbranschen, sjukvården, i väderprognoser och för att analysera kundbeteenden. Särskilt bra fungerar prediktion om det finns *tidsseriedata* att utgå från. Det betyder att mätningar har gjorts många gånger med jämna intervaller, så att man får en sekvens data att söka mönster i.

När människor och AI samarbetar

I stället för att helt överlämna kontrollen till datorn, blir resultatet ofta bäst om människor och datorer löser problem tillsammans. För att det ska fungera behöver AI-system utformas så att de blir lätta och naturliga för människor att interagera med.

Generativ AI kallas AI-modeller som skapar nya data i form av framför allt text, bilder, musik eller datorkod. Ett viktigt exempel är så kallade stora språkmodeller, som den omtalade GPT. Den har kombinerats med en chattfunktion för att skapa tjänsten ChatGPT som gör det möjligt att ha ett människoliknande ”samtal” med språkmodellen. Grunden för tekniken är kraftfulla neurala nätverk (se s. 11) som har tränats på mycket stora mängder språkdata.

Med hjälp av miljarder parametrar kan sådana här språkmodeller tolka text och utifrån tolkningen både komplettera texter och skapa helt nya svar som är mycket lika de som människor ger. Systemen designas och tränas att försöka ge svar som passar i sammanhanget och de kan dessutom göra om och försöka förbättras, om användaren visar att det första svaret inte blev tillräckligt bra.

ChatGPT och flera andra exempel på generativ AI, som Midjourney och Dall-E, blev tillgängliga för allmänheten helt nyligen, men redan använder många miljoner människor programmen för att översätta texter, sammanfatta och analysera. Användare skapar alltifrån konst till propaganda, skriver musik, designar, formger, och söker innehåll på nätet i en samtalsliknande form – AI-systemen blir ett slags sökmotorer som man kan prata med.

Cobots kallas ibland robotar som arbetar tillsammans med människor. Den sortens system har funnits sedan 1990-talet, ofta med målet att avlasta människan från tunga eller repetitiva arbetsuppgifter. Nu börjar AI införas i cobots, vilket skapar robotar som styrs på andra sätt än genom förprogrammering. Med hjälp av tryck- och bildsensorer kan coboten ta emot signaler utifrån och med AI kan den tolka människors språk, gester och fysiska kommandon. En AI-utrustad cobot kan också kontinuerligt ta in ny information och på så vis utvecklas till att fatta allt bättre beslut. Bland annat finns forskningsprojekt om AI-cobots i fysioterapi, där de ska bidra till att göra behandlingen mer effektiv och flexibel.

Samtal direkt med mjukvaran

AI gör det möjligt att i samtalsform, på ett sätt som är naturligt för oss människor, interagera med mjukvara. Vi samspelar alltså med AI på samma sätt som vi skulle göra med en annan människa, utan att behöva förstå modellerna eller tekniken på djupet. Det gör att människor utan stor teknisk kunskap kan styra avancerad teknik på ett sätt som tidigare varit omöjligt. För att exempelvis programmera behövde man tidigare kunna skriva datorkod – nu är det möjligt att istället förklara för en AI vad man vill att ett blivande datorprogram ska göra, så hjälper AI till att skriva koden.

Användningen av AI har snabbt börjat påverka utbildningssystemet i de delar av världen där så gott som alla studenter har tillgång till internet och egen dator. AI-system kan tränas till att bli handledare och guider i lärandet, och helt nya pedagogiska möjligheter uppstår. Ett AI-verktyg kan kommunicera med eleven

FOTO: GUMPAPA/ADOBE STOCK

FOTO: WAVEBREAKMEDIAMICRO/ADOBE STOCK

Det finns många sammanhang där människor och AI arbetar tillsammans. I forskning kan AI bland annat ingå i robotar som avlastar människor, och i undervisning kan AI komplettera läraren genom att skapa övningar som anpassas till elevernas nivå.

och anpassa uppgifternas svårighetsgrad efter behoven. I ett projekt i Norrköping utvecklas ett system som visualiserar molekylmodeller i 3D, samtidigt som användaren kan ställa muntliga frågor och genast få talade svar från systemet. I framtiden kommer vi säkert att få se allt fler system som är byggda för att stimulera flera sinnen och svara på flera sorters signaler.

Nya utmaningar för kunskapssamhället

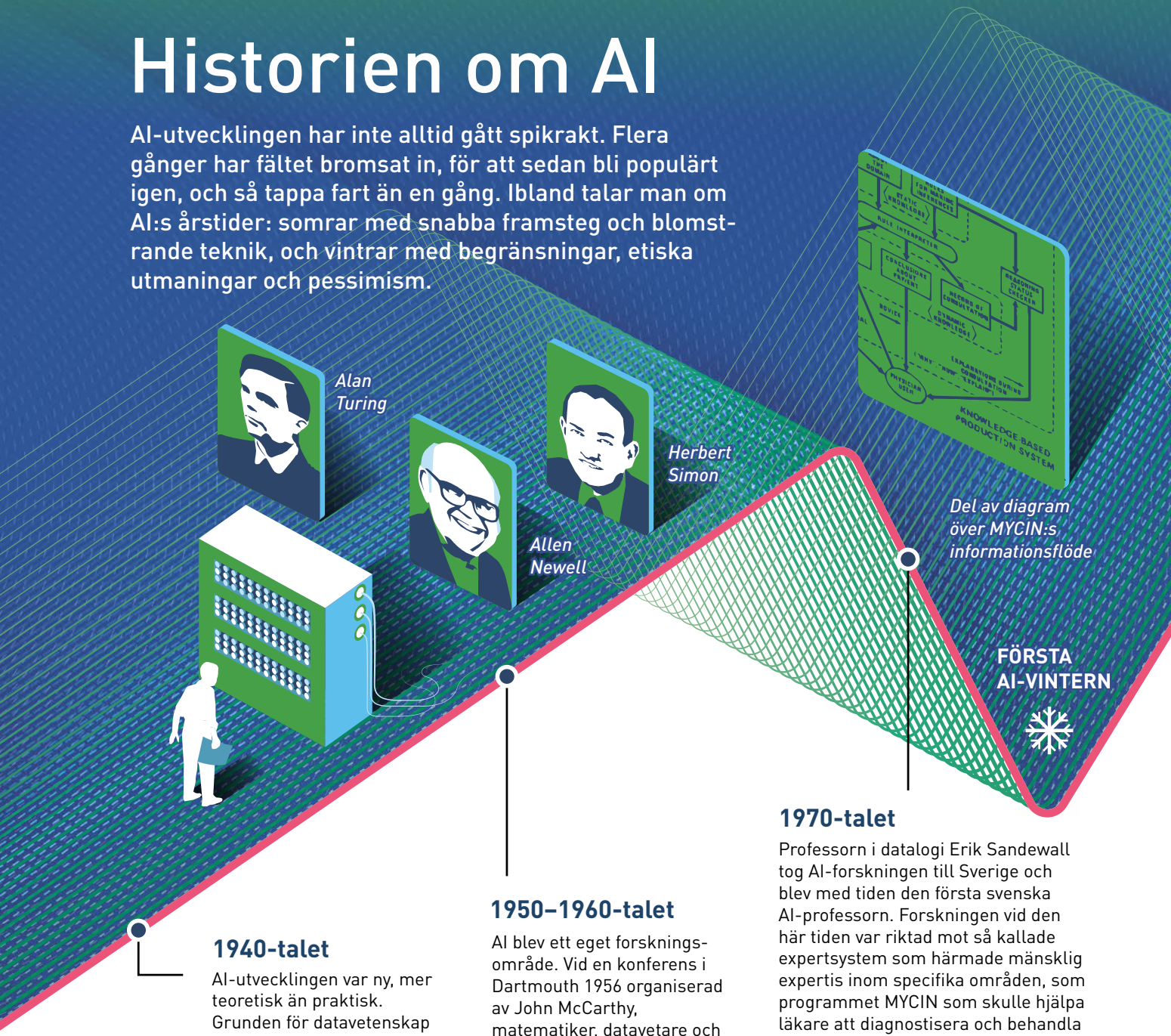
Samtidigt skapar AI nya utmaningar. Den mest uppenbara är risken för fusk. Det blir svårt för läraren att bedöma skriftliga uppgifter om det inte syns ifall eleven skrivit texten själv eller om det är gjort helt eller delvis av en AI.

På sikt behöver även andra risker hanteras. Hur kommer det att påverka vårt lärande om AI sammanfattar, analyserar, läser upp texter åt oss? Lär vi oss lika bra då, bara ännu snabbare än förut? Eller lär vi oss sämre, eftersom det inte är vi som gör själva lärande-jobbet? I dag försöker skolor och universitet hitta sätt att ta sig an de här frågorna. Det är tydligt att det finns ett behov av nya regler och riktlinjer för AI i utbildning.

Att AI kan komma till stor nytta i forskningen är tydligt, men inte heller här är frågan helt okomplicerad. Just nu debatteras bland annat om forskare ska få skriva ansökningar om forskningsmedel, eller till och med vetenskapliga artiklar, med hjälp av AI. Det finns också mer djupgående utmaningar. En mänsklig forskare kan beskriva varje steg på vägen mot ett nytt resultat. Inom vetenskapen är det till och med ett krav att man redovisar sin process så att andra forskare kan kontrollera den och upprepa försöken. AI-system däremot kan inte alltid förklara sina slutsatser. Många gånger är systemen konstruerade så att man matar in en stor mängd data och får ut ett resultat som en människa inte kunde ha tagit fram. Hur vet man då att det stämmer? Och även när man kan slå fast att resultatet stämmer – är det lika värdefullt om man inte kan förstå hur det blev till, och därmed inte kan lära sig av själva processen?

Historien om AI

AI-utvecklingen har inte alltid gått spikrakt. Flera gånger har fältet bromsat in, för att sedan bli populärt igen, och så tappa fart än en gång. Ibland talar man om AI:s årstider: somrar med snabba framsteg och blomstrande teknik, och vintrar med begränsningar, etiska utmaningar och pessimism.



1940-talet

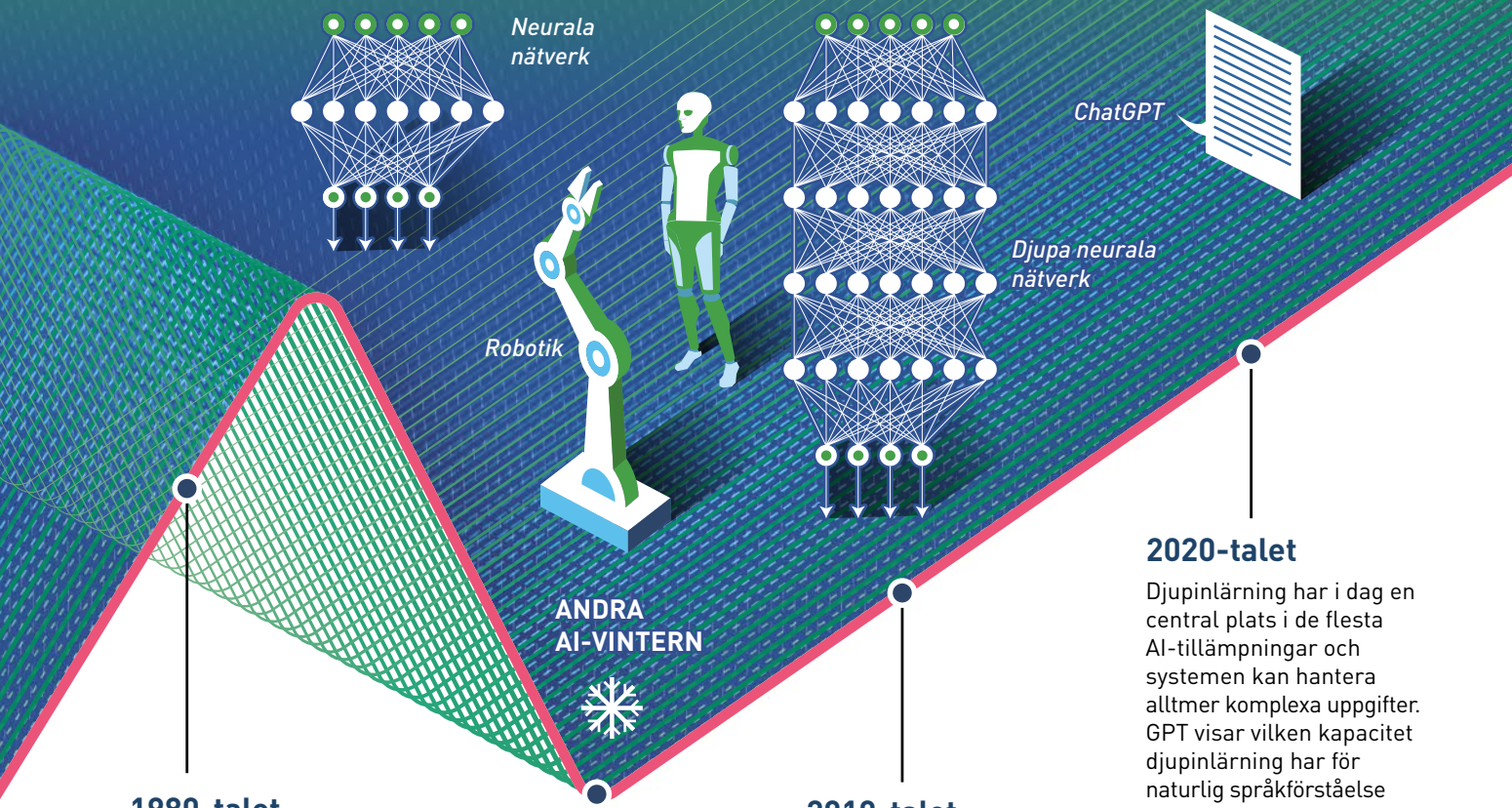
AI-utvecklingen var ny, mer teoretisk än praktisk. Grunden för datavetenskap och maskinintelligens lades av bland andra matematikerna Alan Turing och John von Neumann. Redan 1936 presenterade Turing idén om en universell maskin som kunde räkna som en människa. Under andra världskriget arbetade han med kodknäckningsmaskiner som hade inslag av tidig AI.

1950–1960-talet

AI blev ett eget forskningsområde. Vid en konferens i Dartmouth 1956 organiserad av John McCarthy, matematiker, datavetare och kognitionsvetare, myntades termen artificiell intelligens. Snart kom de första AI-drivna spelen. Datavetaren och psykologen Allen Newell och samhällsvetaren Herbert Simon utvecklade det första AI-programmet, Logic Theorist, som bevisade matematiska satser. Optimismen var stor men de praktiska framstegen små, eftersom datorernas kraft var så begränsad.

1970-talet

Professorn i datalogi Erik Sandewall tog AI-forskningen till Sverige och blev med tiden den första svenska AI-professorn. Forskningen vid den här tiden var riktad mot så kallade expertsystem som härmade mänsklig expertis inom specifika områden, som programmet MYCIN som skulle hjälpa läkare att diagnostisera och behandla svåra bakterieinfektioner (det testades men infördes aldrig). Men vid mitten av decenniet fick det globala AI-fältet problem. Förväntningarna var för stora och många projekt gav inte de utlovade resultaten. Andra hälften av 1970-talet brukar kallas för den första AI-vintern. Finansieringen och det allmänna intresset minskade.



1980-talet

Återhämtning från ”vintern”. Forskarna hade tvingats omvärdera sina metoder, hårdvaran och algoritmerna blev bättre och förväntningarna mer realistiska. Maskininlärning och neurala nätverk fick fotfäste, och företag investerade stort i expertsystem inom bland annat medicin och på finansmarknaden. Men tekniska utmaningar och hård konkurrens sänkte marknaden, och under den senare delen av decenniet gick många AI-bolag i konkurs. Systemen visade sig dyra att underhålla och uppdatera, och gjorde dessutom för allvarliga misstag. Åren omkring skiftet mellan 1980- och 90-talet räknas som den andra AI-vintern.

1990-talet –2000-talet

På 1990-talet började maskininlärningsrevolutionen. Det skapades allt fler datadrivna och statistiska metoder och de nya teknikerna började användas inom språkbehandling, datorseende och robotik. Forskare skapade fristående program som kunde interagera med varandra och med människor. Det la grunden för modernare AI-system, som virtuella assistenter som klarade av sådant som att svara på frågor eller boka biljetter, och rekommendationsalgoritmer som används i sökmotorer och strömningstjänster för musik och film.

2010-talet

Djupinlärning, en typ av maskininlärning, slog igenom och så kallade djupa neurala nätverk förbättrade AI:s kapacitet betydligt. När stora teknikföretag satsade kraftigt på AI-forskning ökade tempot ytterligare. Ny hårdvara som grafikprocessorer (GPU:er) spelade en avgörande roll för att träna djupinlärningsmodeller. Dessutom fanns nu stora mängder digitala och maskinläsbara data att träna på. AI började användas i självkörande fordon, inom vården och i digitala assistenter som Siri och Alexa. Arbetssättet i hela industrier började påverkas och AI blev del av allt fler människors vardagsliv.

2020-talet

Djupinlärning har i dag en central plats i de flesta AI-tillämpningar och systemen kan hantera alltmer komplexa uppgifter. GPT visar vilken kapacitet djupinlärning har för naturlig språkförståelse och med generativ AI skapas text, bilder och musik. Under covid-19-pandemin bidrog AI både till att nya vaccin kunde designas så snabbt, och till att tillverkningen kunde skalas upp på ett effektivt sätt. Samtidigt har etiska frågor och behovet av AI-reglering börjat diskuteras på bred front.

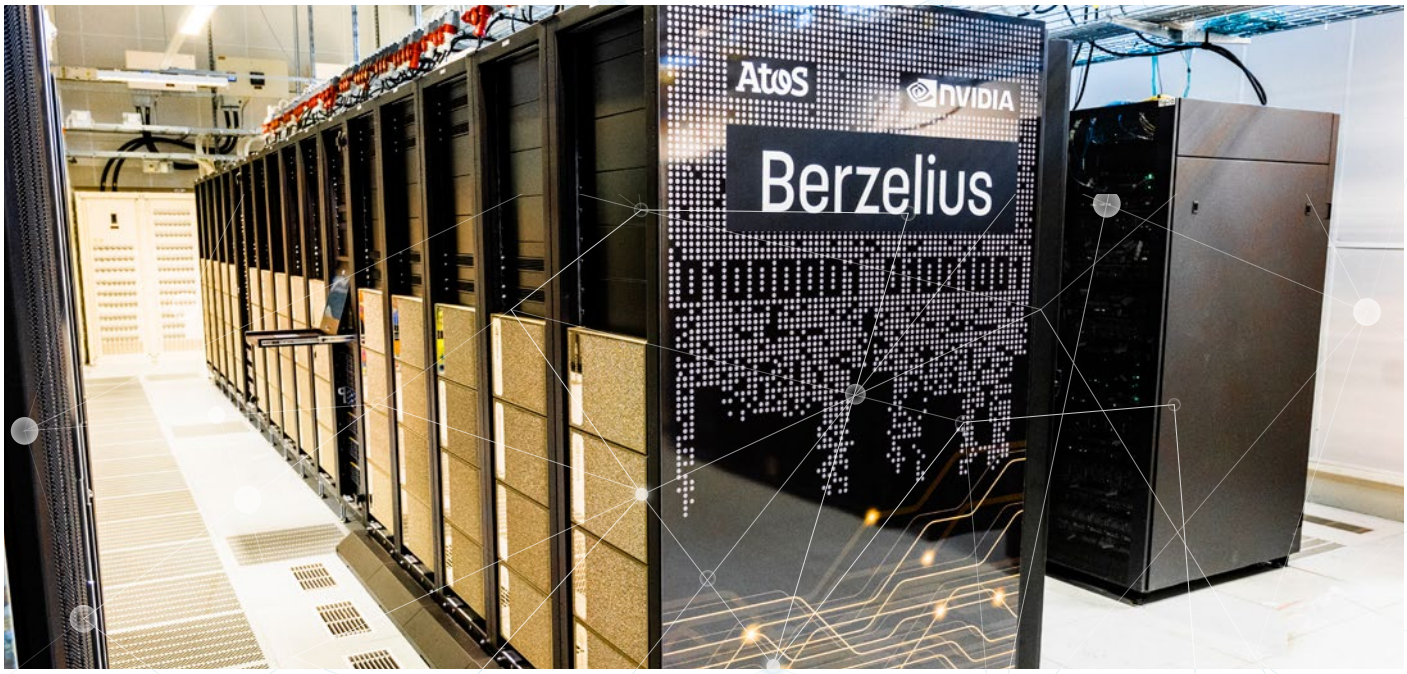


FOTO: THOR BALKHED/LINKÖPINGS UNIVERSITET

2021 togs AI-superdatorn Berzelius i drift vid Nationellt Superdatorcentrum, Linköpings universitet. Den har använts för att träna en svensk språkmodell kallad GPT-SW3 och ingår i forskningsprojekt om bland annat proteinveckning och bildanalys.

Tekniken bakom AI

AI bygger på en mängd metoder och tekniker. Flera av dem kombineras när systemen blir större och mer komplexa. Det finns också flera sätt att träna AI som kräver olika mycket arbete och passar för olika typer av uppgifter.

Klassiska AI-metoder

Det som kallas för *klassisk AI* eller *symbolisk AI* är en samling metoder som använder regler och logik för att hantera och presentera data. Vi människor använder ofta i grunden enkla metoder för att lösa problem, både i spel och lekar – som när vi löser ett sudoku – och i arbetet, när vi vet vad som behöver göras när en viss sak händer. Sökalgoritmer är ett exempel på klassisk AI, liksom planeringssystem som sätter samman en sekvens av åtgärder för att nå ett visst mål.

Regelbaserad AI är en särskild form av klassisk AI. Då programmeras systemet till att använda uttryckliga regler för olika situationer, för att härma en mänsklig expert. Systemen kodas i form av ”om... så...”-regler, alltså att ett visst villkor utlöser en viss förutbestämd handling. Sådana system används ofta som beslutsstöd. De guidar den mänskliga användaren så att den inte behöver hålla alla fakta och regler i huvudet utan kan följa systemets väg, som från början har skapats med mänsklig expertkunskap.

Maskininlärning

Ordet maskininlärning syftar på metoder för att lära algoritmer relationen mellan data som matas in och vilken data som ska matas ut. Inlärningen sker genom att systemet tränas, med inspiration från hur vi människor lär oss. I vårt fall är det oftast sinnena som ger oss indata som vi använder för att skapa en förståelse av vår omgivning och situationen omkring oss. Genom träning lär vi oss hur vi ska använda indata för att fatta beslut och agera. Besluten och handlingarna kan man kalla vår utdata.

För ett AI-system kan indata vara exempelvis bilder, ljud- och filmupptagningar eller digitala dokument,

som data ur elektroniska patientjournaler. Utdata kan vara att systemet klassificerar indata och sätter etiketter, *labels*, på dem. Det kan handla om att slå fast att en bild föreställer en katt eller en väg, eller att ljudströmmen kommer från en viss låt, eller vilken sjukdom som symtomen och provsvaren i patientjournalen tyder på.

Maskininlärning kan också användas för att gissa hur ett skeende ska utveckla sig utifrån historiska data, till exempel i väderprognoser, och till att bearbeta och förenkla komplexa data som sedan kan visualiseras så att de blir lättare för människor att förstå.

De tidigaste typerna av maskininlärning användes för att lösa specifika och väldefinierade problem, som att upptäcka ansikten i bilder eller urskilja vilken e-post som är skräppost. Metoderna byggde på fördefinierade matematiska uttryck som algoritmen sökte efter i indata. Det kunde handla om statistik över vissa ord i e-postmeddelanden, eller den geometriska relationen mellan ögon, näsa och mun för att identifiera ansikten. Det var upp till programmeraren att definiera vad som var relevant för uppgiften.

Nu har mängden data växt och problemen blivit mer komplexa. Det fungerar inte längre att låta människor designa matematiska uttryck. Istället används *representationsinlärning*. Det betyder att maskininlärningsalgoritmer tränas att automatiskt upptäcka relevanta mönster i indata. Träningen kan göras direkt på obearbetad rådata och algoritmerna hittar samband och mönster som är omöjliga för människan att se. Till exempel kan människor lätt definiera visuella koncept som är relevanta för att beskriva olika ansikten och objekt, men har mycket svårare för att specificera vilka inslag som är mest relevanta i ljuddata eller GPS-information, eller i mätdata från sensorer.

Neurala nätverk och djupinlärning

Artificiella neurala nätverk har blivit en mycket viktig del av maskininlärningen. Det är datormodeller som i viss mån påminner om strukturen och funktionen hos människans hjärna. Nätverken kan formges olika, men består alltid av sammanlänkade noder. Noderna är organiserade i lager på ett sätt som gör det möjligt för dem att behandla information och användas till uppgifter som att känna igen mönster och fatta beslut.

Tre träningsmetoder

Det finns flera typer av maskininlärningsmetoder.
De skiljer sig åt bland annat genom hur systemen tränas.

- **Övervakad inlärning** innebär att indata under träningen är märkt, så att det framgår exempelvis om bilden föreställer en katt eller hund. En nackdel med metoden är att den kräver mycket arbete av människor, särskilt om det är stora mängder data som ska märkas (*annoteras*).
- **Oövervakad inlärning** bygger på omärkt data. Systemet definierar självt kategorierna. På så vis kan dolda strukturer eller relationer i data visa sig, samband som människor kanske inte hade kunnat se själva.
- **Förstärkningsinlärning** är när systemet interagerar med omgivningen under träningen. Systemet har tillgång till ett slags miljö och får automatiskt efter varje handling en återkoppling som talar om ifall handlingen var framgångsrik eller inte. Målet är att systemet ska utveckla en strategi som maximerar den positiva återkopplingen.

Så kallad övervakad maskininlärning kräver att människor märker (annoterar) indata.

FOTO: NACHO KAMENOV & HUMANS IN THE LOOP / BETTER IMAGES OF AI / DATA ANNOTATORS LABELING DATA / CC-BY 4.0

En enkel typ av nod i de neurala nätverken är något som kallas för *perceptron*, en konstruktion inspirerad av modeller av hjärnans nervceller. Varje lager av perceptroner bearbetar indata från det föregående lagret och producerar utdata, som blir indata för nästa lager. Kopplingarna mellan perceptronerna fungerar ungefär som neuronerna; de som arbetar med samma sak är sammankopplade, medan de som gör olika saker inte är kopplade till varandra (som att vissa av hjärnans delar är specialiserade på syn, andra på hörsel). Varje koppling kallas för en vikt, och under träningen av nätverket ställs vikternas värde in. De kopplingar som efter analysen av indata rent matematiskt har bidragit mer till slutvärdet får större vikt än de som har bidragit mindre. Genom träning på en stor mängd data optimeras vikterna, och sedan låser man dem när systemet är färdigt att användas. Antalet vikter beror på hur nätverket har designats. I en språkmodell kan det ingå flera miljarder vikter.

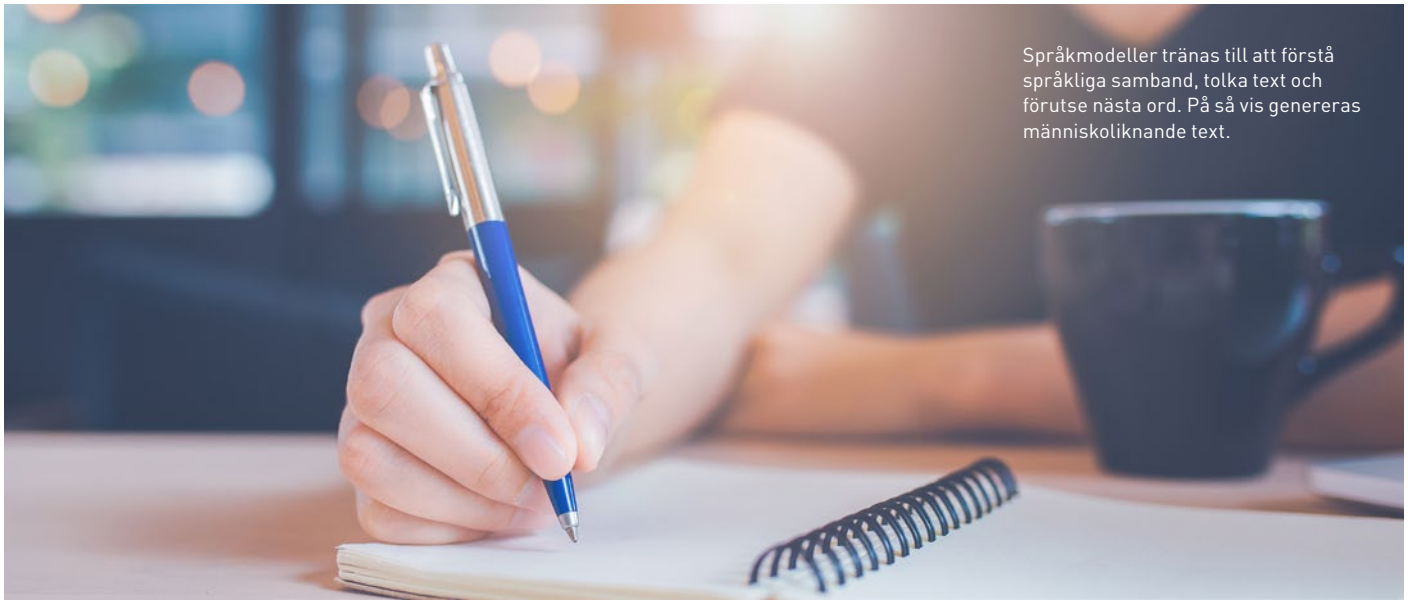
Ett neuralt nätverks struktur brukar kallas dess arkitektur, och i den ingår ofta betydligt fler komponenter än perceptronen. Arkitekturen kan variera i komplexitet och djup, exempelvis antal lager och hur de är sammankopplade.

Neurala nätverk tränas med alla de tre grundmetoderna som nämnts, men också allt oftare genom något man kan kalla *delvis övervakad inlärning*. Det innebär att indata används för att automatiskt skapa både indata och utdata. Ett exempel är träning av språkmodeller som GPT där man börjar med att mata in text från internet. Modellen maskerar sedan slumpvis delar av texten, varpå den tränar sig själv på att fylla i det saknade. På så vis skapar den delvis sin egen indata och processen behöver inte övervakas av människor. Med en arkitektur kallad en *transformer* får nätverket hjälp att hitta samband mellan hur ord oftast används i en mening, och hur meningar oftast används i ett visst sammanhang. När nätverket förstår sambanden kan det tolka text och förutse vilka ord som borde komma härnäst. På så vis genereras människoliknande text.

Djupinlärning är ett begrepp man hör allt oftare. Det handlar helt enkelt om ännu större neurala nätverk med betydligt fler lager, ibland tusentals. Det har lett till nya tekniska genombrott inom bland annat datorseende, och hur väl AI-system kan förstå och bearbeta mänskligt språk.

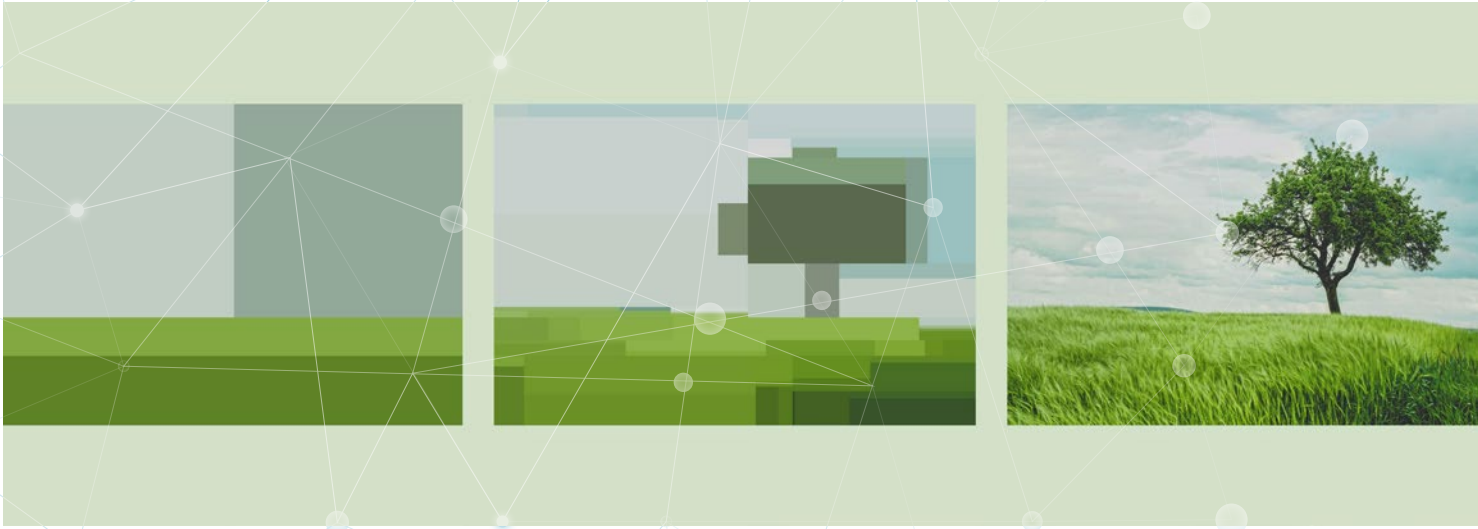
Kategoriserande AI

Kategoriserande AI (*discriminative AI*) kallas modeller fokuserade på att finna gränserna mellan olika kategorier i indata. Modellerna kan bygga på neurala nätverk eller klassiska statistiska metoder, men avsikten är alltid att göra en så bra koppling som möjligt mellan indatans egenskaper och utdatans *labels*, ”etiketter”. Det här är effektiva system om uppgiften är att känna igen mönster eller klassificera något.



Språkmodeller tränas till att förstå språkliga samband, tolka text och förutse nästa ord. På så vis genereras människoliknande text.

FOTO: APICHON_TEE/ADOBE STOCK



RENS DIMMENDAAL & JOHANN SIEMENS / BETTER IMAGES OF AI / DECISION TREE / CC-BY 4.0

En typ av kategoriserande AI är *regressionsmodeller* som kan användas för att bedöma förhållandet mellan en viss variabel man är intresserad av, kanske priset på hus, och ett antal andra variabler som spelar roll i sammanhanget, som läget och storleken. Man kan alltså säga att priset är en kategori som modellen urskiljer, och den kategorin definieras av värdet på de andra variablerna som modellen väger samman.

Generativ AI

En mycket snabbt växande typ av maskininlärning är så kallade generativa modeller. Det är algoritmer utformade för att skapa nya data, till skillnad från de kategoriserande modellerna som bara kan klassificera indata. Generativ AI används för att skapa bilder, text och ljud. Modellerna har sina rötter i den del av matematiken som kallas *probabilistisk modellering*. De lär sig vilken sannolikhetsfördelning som låg bakom indata, omvandlar sannolikheterna (i bilder, ljud eller annat som modellen tränats på) till matematik och använder detta för att skapa ny data som är statistiskt lik den inmatade datan vad gäller mönster, struktur och inbördes relationer.

Stora språkmodeller som GPT (förkortningen står för *Generative Pretrained Transformer*) är exempel på generativ AI. De har gått igenom två utvecklingssteg: förträning och finjustering. Under förträningen lär sig modellen språkmönster och relationer genom att förutsäga nästa ord eller symbol i ett stort dataset. Med symbol menas den minsta meningsfulla enheten i text i naturligt språk, kanske ett ord, en del av ett ord, en bokstav eller ett skiljetecken. Under finjusteringen begränsar man sedan modellens fokus så att den blir mer relevant för ett visst sammanhang. I slutänden blir språkmodellen expert på att förutsäga vilket nästa ord, eller vilken nästa symbol, bör vara.

Multimodal AI

Människor uppfattar världen genom olika sinnen: syn, hörsel, känsel, smak och lukt. Ibland kallar man det här för olika *modaliteter*, som tillsammans skapar en djupare förståelse av miljön än vad varje enskilt sinne gör. Multimodal AI syftar på AI som kan bearbeta och förstå information från flera modaliteter, eller källor.

Ett exempel är AI-system som kan skapa bilder med hjälp av text. Ofta bygger de på att en språkmodell omvandlar texten till en prompt, ett slags instruktion, till en bildgenerator. Bildgeneratorn bygger i många fall faktiskt på konceptet bild*försämring* – nätverket har tränats på att förstå hur en bild försämras när den ”blurras”, görs suddigare (vilket man ofta gör för att avidentifiera en person eller en plats). När nätverket förstår hur den processen går till, kan det lära sig att driva den åt andra hållet. Prompten används för att styra skapandet av en ny bild.

Multimodal AI är också viktigt i bland annat självkörande fordon, där information från flera källor behöver kombineras.

Vägen mot framtidens AI

FOTO: AKLIMA/ADOBE STOCK

Efter decennier av forskning har AI tagit steget hela vägen ut till användare på alla nivåer i samhället, från privatpersoner till myndigheter. Hur kommer utvecklingen att se ut framöver? Hur kraftfull kan AI bli, och hur ska vi maximera de goda effekterna och minimera riskerna?

De närmaste åren får vi troligen se AI, stora datamängder och robotteknik kombineras i fler fysiska AI-system. Människor kommer att interagera alltmer med AI, vilket kan minska belastningen på våra egna hjärnor samtidigt som det ökar produktiviteten och gör det möjligt att göra nya typer av vetenskapliga upptäckter. AI kan mycket väl bli en kollega vars kraftfulla algoritmer gör det möjligt för oss att bättre förstå världen omkring oss.

Men de stora framstegen har inte bara väckt jubel, utan också oro. I dag diskuteras hur säkra AI-systemen är och vilka konsekvenserna kan bli av att använda dem. AI arbetar snabbt, självständigt och ofta utan transparens. Ibland vet varken användaren eller systemets utvecklare hur AI har fattat ett visst beslut. När allt fler människor, företag och organisationer börjar dra nytta av tekniken så har också händelser inträffat som visar på svagheter, och hur systemen kan användas för att avsiktligt skada. Ett välkänt exempel är falska nyheter och när så kallade ”deep fakes” utger sig för att vara äkta. Deep fakes är filmer eller ljudklipp där någons ansikte eller röst ändras. Det förekommer också bedrägerier där AI används för att härma någons röst så att människor tror att de blir uppringda av en släkting eller vän som säger att den behöver pengar snabbt.

Även slarvig användning av AI kan ge problem. Stora språkmodeller som ChatGPT kan ge felaktiga svar och presentera påhittad information. Användaren har själv ett ansvar för att granska svaret innan det används eller sprids. Det är inte minst viktigt vid utbildning, där eleverna måste tränas i kritiskt tänkande och fakta-granskning för att rätt bedöma information och källor.

Vissa skevheter och fel har drabbat många människor på en gång. Det har bland annat inträffat när AI har använts i samhällsfunktioner, som vid bedömning av behovet av ekonomiska bidrag. I Nederländerna anklagade skattemyndigheten familjer för att ha fått felaktiga barnbidrag, utifrån riskfaktorer som angavs av ett AI-system. Innan det framkom att systemet inte fungerade hade tusentals människor krävts på pengar. Många tvingades in i fattigdom, och det förekom till och med självmord och fosterhemsplaceringar. Alla fel får naturligtvis inte så katastrofala konsekvenser, men även mindre fel kan drabba individer hårt. AI-system kan spä på orättvisa och diskriminering, till exempel när ansiktsigenkänningssystem varit sämre på att skilja på mörkhyade än ljushyade ansikten, eller när AI-system för jobbrekrytering automatiskt har prioriterat män.

Det är vi som styr AI

Ibland hör man om *etisk AI* eller *ansvarsfull AI*. Det betyder inte att datasystem kan ta ansvar. Istället syftar det på att de som konstruerar eller använder systemen måste ta fullt ansvar för besluten och för sitt val att använda AI. Det måste också vara tydligt när AI används. Styrningen av AI är avgörande för

att tekniken ska utvecklas på ett sätt som gynnar oss alla. Nu diskuteras regler och riktlinjer bland beslutsfattare, forskare och även människor i allmänhet.

Diskussionerna om AI:s effekter bör sträcka sig bortom risken för direkta skador. Vi behöver fråga oss vilka aktiviteter vi vill utföra själva, och vad som händer när vi överlämnar uppgifter till AI. I vilka sammanhang blir resultatet bättre och vi befrias från onödig belastning – och i vilka sammanhang är det viktigt att vi fortsätter att på egen hand skapa, lära eller besluta?

Vad händer med olika yrken när AI kan utföra fler uppgifter, och hur ska vi säkra en rättvis utveckling som inte gör vissa grupper av människor till förlorare? Hur ska vi få miljömässigt hållbar AI, när tekniken i dag kräver stora mängder energi? Hur ska vi hantera upphovsrättsfrågor när generativ AI producerar nytt utifrån material som människor har skapat? Redan har konstnärer, artister och journalister börjat kräva att AI inte gratis ska få tränas på deras material. Just nu finns inga bra regelsystem för det, och mycket av utvecklingsarbetet på AI-området görs av teknikföretag som hemlighåller vilken data de använder och hur de arbetar.

Det är inte ovanligt att människor är rädda för AI. Ofta beskrivs AI som något som händer oss, något vi inte kan kontrollera utan bara försöka lindra konsekvenserna av. Det ger en känsla av maktlöshet. Men AI händer inte oss – det är vi som får AI att hända. AI skapas och designas av människor och det ger oss alternativ och val. Genom ansvarsfull utveckling, tydliga regler och noggranna etiska överväganden kan vi säkra att risken för skador blir så liten som möjligt och att AI används för samhällets bästa.

Artificial Intelligence Act

AI-förordningen, eller ”the AI Act”, är en lagsamling för hela EU – det mest omfattande AI-regelverket i hela världen. De nya lagarna ställer krav på säkerhet, etik och mänskliga rättigheter utifrån fyra riskkategorier:

- **Oacceptabel risk.** Sådan AI förbjuds. Exempel: AI som används för manipulation, utnyttjande och social kontroll. Som de automatiska rankingsystem som bland annat Kina använder för sina medborgare.
- **Hög risk.** De här systemen kommer behöva följa alla förordningens krav. Exempel: AI i transportsystem, utbildningssystem, sjukvård, CV-granskning vid rekrytering, juridisk bevisvärdering.
- **Begränsad risk.** Färre av kraven behöver uppfyllas. Exempel: chattbottar.
- **Minimal risk.** Regleras inte. Exempel: spamfilter och TV-spel.

AI-förordningen kräver bland annat transparens, noggrann testning, att problem rapporteras och att systemen håller en viss nivå av cybersäkerhet. Andra exempel på AI-regler är den amerikanska regeringens *Blueprint for an AI Bill of Rights*, och FN-organet Unicefs vägledande riktlinjer för AI för barn.

AGI och debatten om X-Risk

Artificiell Generell Intelligens (AGI) skulle vara ett AI-system som kan göra allt som vi människor kan, och mer. En del AI-forskare anser att vi är ganska nära detta – andra att det dröjer decennier, eller att AGI aldrig kommer att finnas. Flera stora företag tävlar i dag om att först lyckas utveckla AGI. Nu växer en debatt om så kallad existentiell risk, X-Risk: att AGI utom kontroll skulle kunna hota vår existens. Vad händer om vi skapar AGI, ger den stort inflytande över den fysiska världen och otillräckliga instruktioner? Vissa forskare menar att systemen skulle kunna prioritera andra mål högre än vår trygghet och skada oss för att uppnå de målen.

AI-experter är inte överens om hur stor X-Risk är och när, eller ens om, den blir aktuell. Men de som varnar vill att även den risken tydligt vägs in i AI-regleringar.

KUNGL. VETENSKAPS-AKADEMIEN
THE ROYAL SWEDISH ACADEMY OF SCIENCES

VETENSKAPEN SÄGER ★ NR 6 ★ APRIL 2024

En expertgrupp bestående av ledamöter av Vetenskapsakademien och inbjudna experter har tagit fram materialet till denna skrift som bygger på den samlade vetenskapliga litteraturen. Skriften speglar expertgruppens uppfattning och ska inte ses som ett uttalande eller ställningstagande av Kungl. Vetenskapsakademien:

VIRGINIA DIGNUM, professor i AI, inriktning social och etisk AI, Umeå universitet

FREDRIK HEINTZ, professor i datavetenskap, inriktning AI och autonoma system, Linköpings universitet

DANICA KRAGIC JENSFELT*, professor i datavetenskap, inriktning robotik och autonoma system, KTH

AMY LOUTFI, professor i datavetenskap, inriktning AI och robotik, Örebro universitet

ANDERS YNNERMAN*, professor i vetenskaplig visualisering, Linköpings universitet

* Ledamot av
Kungl. Vetenskapsakademien

VETENSKAPEN SÄGER är en serie populärvetenskapliga skrifter från Kungl. Vetenskapsakademien. Målsättningen är att sprida vetenskapsbaserad information om viktiga och aktuella ämnen till allmänheten, särskilt på områden där forskningen har gjort stora framsteg på senare tid.

Beställ tryckta exemplar av Vetenskapen säger från vetenskapensager@kva.se eller ladda ned i pdf-format på www.kva.se/vetenskapensager.

Vetenskapsredaktör: Lisa Kirsebom och expertgruppen för *Vetenskapen säger – om AI*

Omslagsfoto: Eman/Adobe Stock

Illustrationer:
©Johan Jarnestad/
Kungl. Vetenskapsakademien
©Kicki Ajax/Fräulein Design

Grafisk form: ©Fräulein Design

Vetenskapen säger – om AI
produceras och distribueras genom stöd från Stiftelsen Natur & Kultur.