ChatGPT

# Ordlista över centrala AI-begrepp

## AI i undervisning

**Definition:** Användning av artificiell intelligens för att stödja undervisnings- och lärandeprocesser.

**Beskrivning:** AI i undervisning innebär att AI-teknik tillämpas i utbildningssammanhang för att underlätta lärande och undervisning. Det kan handla om allt från **intelligenta tutorer** som anpassar sig efter elevens kunskapsnivå till **automatiserad bedömning** av prov. Potentiellt kan AI hjälpa till att individanpassa undervisningen, avlasta lärare genom att automatisera rutinuppgifter samt ge elever omedelbar feedback [1]. Samtidigt finns utmaningar och risker, bland annat relaterade till integritet, bias och tillförlitlighet, som behöver hanteras för att AI ska **användas ansvarsfullt och rättvist i skolan** [1]. UNESCO framhåller att AI kan adressera stora utbildningsutmaningar och förnya pedagogiken, men betonar att tekniken måste användas med principer om inklusion och rättvisa i fokus [1].

**Exempel:** *En gymnasielärare kan använda en AI-driven språkassistent i klassrummet för att hjälpa elever träna konversation på ett främmande språk, där AI:n ger direkt feedback på uttal och grammatik. På så sätt får varje elev öva i sin egen takt, medan läraren frigörs för att ge mer individuell stöttning.*

**Källor:** UNESCO (2023) understryker AI:ns potential att *"lösa några av de största utmaningarna i dagens utbildning"* samtidigt som snabba framsteg *"medför flera risker och utmaningar"* som måste hanteras ansvarsfullt [1].

---

## AI-agenter

**Definition:** Programvaru-**agenter** drivna av AI som kan uppfatta sin omgivning, fatta beslut och utföra handlingar självständigt för att uppnå givna mål [2].

**Beskrivning:** En AI-agent är ett **autonomt program** som, likt en digital "agent", **observerar sin miljö**, bearbetar information och agerar utifrån givna uppgifter. AI-agenter utgår från sensoriska input (t.ex. data eller signaler) och använder beslutslogik – ofta med hjälp av *maskininlärning* – för att välja handlingar som för dem närmare sina mål [3]. En **rationell AI-agent** strävar efter optimala beslut baserat på sina perceptioner och erfarenheter [4]. Enkla exempel är **chattbottar** som tar in användarfrågor och ger svar, medan mer avancerade agenter kan planera flera steg framåt utan mänsklig inblandning. Moderna AI-agenter kan ges **hög grad av autonomi**, och kan t.ex. hämta data, samverka med andra system och kontinuerligt lära av resultat för att förbättras [5]. Graden av autonomi bestäms dock av utvecklarna och kan begränsas för säkerhet och pålitlighet [6].

**Exempel:** *En AI-agent i form av en dammsugarrobot navigerar själv genom ett hem: den använder sensorer för att "se" hinder, planerar sin rutt genom rummen och suger upp smuts – allt utan att användaren behöver styra den manuellt.*

**Källor:** AWS beskriver AI-agenter som *"programvaror som kan interagera med sin omgivning, samla in data och använda dessa för att självständigt utföra uppgifter som möter förutbestämda mål"* [2]. Oracle

framhåller att avancerade agenter kan utföra **flerstegsuppgifter** och fatta beslut på egen hand, men att *"autonomin som ges till agenter bestäms av de människor som initierar dem"* [5].

---

## AI-assistenter

**Definition:** Digitala assistenter drivna av AI som förstår **naturligt språk** (tal eller text) och utför uppgifter eller ger information åt användare [7].

**Beskrivning:** AI-assistenter (även kallade **virtuella assistenter**) är mjukvaror som utnyttjar **språkförståelse** och andra AI-tekniker för att interagera med människor på ett naturligt sätt. De kan tolka **röstkommandon eller textfrågor** och utföra åtgärder såsom att besvara frågor, hantera kalendrar, skicka meddelanden, ge väderrapporter eller styra smarta hem-enheter. Kända exempel är **Amazon Alexa**, **Apple Siri** och **Google Assistent**, vilka alla använder taligenkänning och språkgenerering för att kommunicera med användaren. En AI-assistent har ofta tillgång till internet och olika tjänster för att utföra uppdrag – till exempel boka möten, läsa upp nyheter eller spela musik. Tekniken bygger på avancerad **naturlig språkbehandling** (NLP) och maskininlärning, vilket gör att assistenten kan förbättras över tid och anpassa sig efter användarens preferenser.

**Exempel:** *En lärare kan använda en AI-assistent i form av en smart högtalare i klassrummet. Läraren säger: "Hej assistent, sök fram fakta om Mars." Assistenten tolkar uppmaningen och svarar högt med relevant information hämtad från nätet, vilket sparar tid och gör lektionen mer interaktiv.*

**Källor:** TechTarget definierar en AI-assistent som mjukvara som *"använder artificiell intelligens för att förstå röstkommandon i naturligt språk och utföra uppgifter åt användaren"* [7]. Populära AI-assistenter som **Alexa, Siri och Google Assistant** klarar uppgifter motsvarande en personlig sekreterare – till exempel *"ta diktamen, läsa upp meddelanden, slå upp telefonnummer, schemalägga samtal och påminnelser"* [8].

---

## AI-litteracitet

**Definition:** Förmågan hos individer att förstå, använda och kritiskt förhålla sig till AI-teknik och dess effekter på samhället [9].

**Beskrivning:** Begreppet AI-litteracitet (artificiell intelligens-läskunnighet) handlar om **kunskaper och färdigheter** som gör det möjligt för människor – i synnerhet icke-experter – att navigera en värld där AI-system är vanliga. Det innefattar att förstå **hur AI-system fungerar**, vilka möjligheter och begränsningar de har, samt att kunna tolka AI-genererade resultat på ett kritiskt sätt. En AI-litterat person vet exempelvis att en AI-modell lär sig från data och att dess beslut kan vara felaktiga eller partiska beroende på träningsdatan. AI-litteracitet betonar även att kunna **använda AI-verktyg säkert och etiskt**, till exempel att vara medveten om integritetsfrågor och att AI behöver övervakning. På samhällsnivå anses AI-litteracitet bli allt viktigare – många menar att skolor bör lära ut grunderna i AI, så att framtidens medborgare är rustade att både utnyttja AI-teknik och förstå riskerna [9] [10].

**Exempel:** *En lärarstudent med god AI-litteracitet kan använda verktyg som AI-drivna språkmodeller i sina studier – till exempel för idéer till lektionsplanering – men gör det med förstånd. Hen förstår att AI:ns förslag kan behöva granskas kritiskt och anpassas, och att källor måste dubbelkontrolleras.*

**Källor:** Wikipedia definierar AI-litteracitet som *"förmågan att förstå, använda, övervaka och kritiskt reflektera över AI-applikationer"* [9]. Vidare beskrivs det som *"en uppsättning kompetenser som gör det möjligt för individer att kritiskt utvärdera AI-teknologier; kommunicera och samarbeta effektivt med AI; och använda AI som ett verktyg"* [10].

---

## AI-organisationer

**Definition:** Företag, forskningsinstitut eller ideella sammanslutningar som fokuserar på utveckling av AI-teknik, AI-policy eller AI-forskning i syfte att främja fältet och säkerställa dess ansvarsfulla användning.

**Beskrivning:** AI-organisationer kan ta många former: **kommersiella aktörer** (t.ex. teknologiföretag och startup-bolag med AI-produkter), **akademiska laboratorier** vid universitet som forskar i AI, samt **ideella organisationer och samarbeten** som arbetar med AI-etik och riktlinjer. Dessa organisationer driver fram AI-utvecklingen och formar hur tekniken används i samhället. Exempel på inflytelserika AI-organisationer är **OpenAI** – en forskningsorganisation känd för att utveckla bl.a. språkmodellen GPT-4 – samt **DeepMind** (en del av Google) som gjort genombrott inom förstärkningsinlärning. Det finns också samarbetsorgan såsom **Partnership on AI**, som grundades 2016 av flera tech-företag (bl.a. Google, Facebook, Amazon, IBM, Microsoft) för att *"öka allmänhetens förståelse för AI samt ta fram riktlinjer för etik, transparens, rättvisa och tillförlitlighet"* inom AI-forskning [11]. På det globala planet finns initiativ som **Global Partnership on AI (GPAI)** och olika FN-organ som engagerar sig i AI-frågor. **AI Sweden** är ett svenskt exempel – ett nationellt centrum som samlar företag, akademi och offentlig sektor för att påskynda tillämpningen av AI i Sverige. Gemensamt för dessa organisationer är att de formar AI-fältets riktning, antingen genom tekniska framsteg eller genom att sätta upp etiska och regulatoriska ramar.

**Exempel:** *I april 2023 samlades representanter från ledande AI-organisationer vid ett globalt toppmöte för att diskutera AI-säkerhet. Bolag som OpenAI och DeepMind deltog tillsammans med forskare och lagstiftare för att enas om riktlinjer som kan förhindra missbruk av avancerade AI-system.*

**Källor:** The Guardian rapporterade vid bildandet av **Partnership on AI** att alliansen syftar till att *"öka den allmänna förståelsen av [AI]-sektorn, samt ta fram standarder som framtida forskare bör följa"*, med fokus på områden som **etik, rättvisa, transparens, integritet** och samarbete mellan människa och AI [11] [12]. Wikipedia noterar att OpenAI (grundat 2015) har som mål att utveckla *"säker och gynnsam"* AI (specifikt generell AI) för mänskligheten [13], vilket exemplifierar hur en AI-organisation kan drivas med ett uttalat etiskt syfte.

---

## AI-rättigheter

**Definition:** Principer och lagförslag som rör **rättigheter i en AI-driven värld** – dels människors rättigheter gentemot AI-system (t.ex. rätt till insyn, rättvisa och integritet när AI fattar beslut om dem), dels den teoretiska frågan om avancerade AI någon gång bör tillerkännas egna "rättigheter".

**Beskrivning:** Diskussionen om AI-rättigheter har två huvudsakliga dimensioner. För det första handlar det om att skydda **individers rättigheter i AI-systemens era**. Exempelvis har USA tagit fram en *AI Bill of Rights* – riktlinjer för att säkerställa att AI-system är säkra, att beslut kan förklaras för de berörda, att inga grupper diskrimineras och att privat data skyddas [14] [15]. Inom EU diskuteras liknande principer i AI-förordningen (AI Act) under utarbetande. Syftet är att medborgare ska ha rättigheter som hindrar skada från AI, exempelvis rätten att få en mänsklig översyn av automatiserade beslut som påverkar

dem. För det andra förs en mer futuristisk och filosofisk diskussion: *Om* AI når medvetande eller personliknande förmågor, bör sådana AI då ha **egna rättigheter** (t.ex. att inte godtyckligt stängas av eller missbrukas)? Denna fråga är i dagsläget teoretisk, men har uppmärksammats i sci-fi och av enstaka forskare. Några AI-forskare (t.ex. på företaget **Anthropic**) har nyligen lyft fram begreppet *"AI-välfärd"* – de undersöker om framtida **självmedvetna AI** skulle kunna uppleva något som kräver moraliskt hänsynstagande [16]. I nuläget är konsensus att dagens AI saknar medvetande och därför inte har rättigheter, men debatten fungerar som tankelek om en framtid med eventuella *"elektroniska personer"*.

**Exempel:** *Vita Husets* Blueprint for an AI Bill of Rights *(2022) listar fem grundprinciper, bland annat att AI-system ska vara säkra och fungera förutsägbart, att användare har* "rätt till förklaring" *på viktiga beslut som fattas av algoritmer, och att individen ska kunna välja bort automatiserade system i kritiska sammanhang* [14] [15].

**Källor:** En artikel i *Computer Sweden* beskriver hur Vita huset lanserat en **AI Bill of Rights** med riktlinjer för ansvarsfull AI-utveckling, där grundtanken är *"att slå vakt om konsumenternas rättigheter och skydda människor från potentiella hot från AI"* [14]. Samtidigt rapporterar *Neuron Expert* att forskare som Sam Altman och företag som Anthropic börjat utforska frågor om *"AI-rättigheter"* för avancerade, eventuellt medvetna, framtida system – vilket utmanar oss att tänka nytt kring styrning och etik för kraftfull AI [17].

---

## AI-rättsfall

**Definition:** Juridiska fall och rättsprocesser som involverar AI – antingen där AI-system är orsak eller föremål för tvist, eller där rättsliga frågor uppstår kring AI:s status, ansvar eller konsekvenser.

**Beskrivning:** I takt med att AI-system används inom fler områden har de också blivit föremål för juridiska prövningar. **AI-rättsfall** kan gälla en rad frågor. Vissa fall rör **ansvar och skadestånd**: till exempel om en självkörande bil orsakar en olycka – vem bär det legala ansvaret, tillverkaren eller användaren? Andra fall handlar om **immateriella rättigheter**: kan en AI stå som uppfinnare i en patentansökan, eller upphovsman till ett verk? Hittills har domstolar i flera länder (USA, EU, Storbritannien) fastslagit att endast människor kan vara patent-innehavare, efter uppmärksammade fall där en AI kallad *DABUS* listades som uppfinnare [18] [19]. Ytterligare en kategori gäller **diskriminering och rättvisa**: om en algoritm för utlåning eller rekrytering visat sig missgynna en viss grupp, kan det leda till rättsprocess för brott mot diskrimineringslagar. Vi ser också rättsfall kring **upphovsrätt** – exempelvis har konstnärer stämt AI-företag för att deras verk använts som träningsdata utan tillstånd. Generativ AI som skapar bilder och texter har väckt frågor om *intrång i upphovsrätt* och sådana fall är nu under prövning i flera länder [20] [21]. Domstolar börjar även hantera hur **bevisvärde** påverkas av AI (t.ex. deepfakes) och hur man autentiserar digitalt material. Överlag håller juridiken på att komma ikapp AI-tekniken, och de tidiga rättsfallen sätter prejudikat för framtiden [22].

**Exempel:** *Ett uppmärksammat AI-relaterat rättsfall är Huang vs. Tesla (2018), där familjen till en förolyckad förare stämde biltillverkaren och hävdade att dess autopilot (ett AI-system) var defekt och orsakade dödsolyckan. Under processen argumenterade Tesla att förarens egna handlingar och eventuella "deepfake"-videor av vd:ns uttalanden försvårade bedömningen, vilket tvingade domstolen att ta ställning till AI-systemens roll och bevisvärde på ett nytt sätt* [23] [24].

**Källor:** Thomson Reuters konstaterar att domstolar *"alltmer brottas med grundfrågorna om hur artificiell intelligens ska behandlas annorlunda än människor och hur den påverkar gällande lagar"* [22]. En

genomgång av generativa AI-mål visar en skarp ökning av upphovsrättsprocesser, t.ex. fallet **Andersen v. Stability AI** där konstnärer hävdar att AI-modeller olovligen tränats på deras verk [20].

---

## AI-säkerhet

**Definition:** Tvärvetenskapligt fält som fokuserar på att förebygga att AI-system orsakar oavsiktliga skador, missbruk eller andra skadliga konsekvenser – och att se till att AI är **robust, pålitlig och följer mänskliga värderingar** [25].

**Beskrivning:** AI-säkerhet innefattar både **tekniska åtgärder** och **policyåtgärder** för att minimera risker med AI. Tekniskt handlar det om att utveckla metoder för att göra AI-system **förklarbara, testbara och motståndskraftiga** mot fel. Det gäller att undvika *olyckor* (t.ex. att en autonom bil missidentifierar ett objekt och orsakar krasch) och *missbruk* (t.ex. att någon manipulerar en AI för skadliga syften). Forskning inom AI-säkerhet adresserar teman som **AI-alignment** (hur man får avancerade AI att följa mänskliga etiska värderingar), **motverkande av bias** (så att AI-beslut inte är orättvisa) samt **robusthet** (att AI presterar pålitligt även under ovanliga förhållanden) [25] [26]. På senare år har även **existentiella risker** fått uppmärksamhet inom AI-säkerhet – farhågor att mycket kraftfull AI i framtiden kan utgöra en global katastrofrisk om den inte kontrolleras [26] [27]. Utöver tekniken arbetar AI-säkerhet även med att ta fram **normer, lagar och uppförandekoder**. År 2023 såg vi ökat fokus: ledande AI-företrädare utfärdade gemensamma uttalanden om behovet av säkerhetsåtgärder, och länder som USA och Storbritannien etablerade särskilda AI-säkerhetsinstitut [28].

**Exempel:** *Inom AI-säkerhet kan en åtgärd vara att stress-testa en språkmodell för att se om den kan luras att ge farliga instruktioner. Genom att identifiera sådana* "failure modes" *tidigt kan utvecklarna justera modellen eller lägga in skyddsmekanismer, så att den inte kan utnyttjas för skadliga ändamål.*

**Källor:** Wikipedia beskriver AI-säkerhet som ett fält fokuserat på att *"förhindra olyckor, missbruk eller andra skadliga konsekvenser som kan uppstå från AI-system"*, och noterar att det omfattar både **maskinetik** och **AI-alignment** för att säkerställa att AI är moraliskt och nyttigt [25]. Forskare lyfter aktuella risker som *"kritiskt systemfel, bias och AI-övervakning"* samt nya risker såsom *"teknologisk arbetslöshet, digital manipulation, vapenisering och cyberattacker"*, liksom spekulativa risker med framtida AGI [27].

---

## AI-transparens

**Definition:** Strävan efter att göra AI-systemens beslutsgång och uppbyggnad **begripliga och insynsvänliga** för människor, så att man kan förstå *varför* en AI kom fram till ett visst resultat.

**Beskrivning:** AI-transparens handlar om att minska *"black box"*-fenomenet hos komplexa AI-modeller. Många moderna AI – exempelvis djupa neurala nätverk – fungerar som **svarta lådor**, där det kan vara svårt för användare (eller till och med utvecklare) att veta exakt hur input leder till viss output [29]. Transparens innebär att AI-system ska ge förklaringar eller på annat sätt låta sig inspekteras. Detta kan uppnås genom **Explainable AI (XAI)**-metoder som bryter ner modellens beslut: t.ex. markera vilka faktorer som påverkade beslutet mest. Transparens är viktig för att bygga **förtroende** – en användare eller en ansvarig myndighet vill kunna verifiera att AI-systemet inte diskriminerar eller gör stora fel. Det är också centralt för **ansvarsutkrävande**: om en AI orsakar en skada måste man kunna utreda vad som gick fel. Regulativa initiativ som EU:s kommande AI Act förväntas ställa krav på viss transparens, särskilt för högrisk-AI-system [30]. Samtidigt är ökad transparens en teknisk utmaning, eftersom komplexa

modeller ofta saknar enkla förklaringsmodeller. Här pågår mycket forskning – exempelvis utvecklas verktyg som visualiserar vilka neuroner i ett nätverk som aktiveras, eller förenklade surrogatmodeller som approximerar AI:ns logik på ett begripligt sätt.

**Exempel:** *En bank som använder ett AI-system för låneansökningar måste kunna förklara för en kund varför hen fick avslag. Genom en transparenslösning kan AI-systemet generera en lista av de viktigaste faktorerna – t.ex. "din inkomst var under tröskel X" eller "du har nyligen missat flera betalningar" – så att kunden får en begriplig motivering istället för ett mystiskt "avslag utan förklaring".*

**Källor:** AIUC noterar att AI-system *"ofta fungerar som 'svarta lådor', vilket kan göra det svårt för användare att känna tillit till resultaten"*, och att **Explainable AI** utvecklats specifikt för att *"ge insyn i hur AI fattar sina beslut"* [29]. Förordningar som EU AI Act väntas kräva att företag kan *"förklara hur deras AI-system fungerar, anpassat efter systemets risknivå"* [30], vilket understryker transparensens växande roll.

---

## Algoritm

**Definition:** En **algoritm** är en **steg-för-steg-procedur** eller uppsättning instruktioner för att lösa en uppgift eller ett problem [31].

**Beskrivning:** Inom datavetenskap och AI utgör algoritmer själva **recepten för beräkningar**. En algoritm tar emot viss input (data eller förutsättningar) och anger exakt vilka operationer som ska utföras, i vilken ordning, för att producera önskat output. Algoritmer kan vara så enkla som en serie instruktioner för att sortera en lista med tal, eller extremt komplexa som i AI-sammanhang (t.ex. algoritmen en neuralt nätverk följer för att justera sina vikter under träning). I AI används algoritmer både för **träning** (lärande) – t.ex. gradientdescent-algoritmen som optimerar en modell på data – och för **inferens** (själva beslutsfattandet) – t.ex. en beslutsträdsalgoritm som traverserar ett träd för att ge ett svar. Kännetecknande för algoritmer är att de är **deterministiska** i beskrivning (varje steg är väldefinierat), även om vissa algoritmer (som genetiska algoritmer) innehåller slumpmoment. En bra algoritm är korrekt (ger rätt resultat) och effektiv (använder rimlig tid/minne). Begreppet är grundläggande: all mjukvara består i grunden av algoritmer.

**Exempel:** *En simpel algoritm i vardagen är ett matlagningsrecept: ingredienser utgör input, stegen (blanda, värm i ugnen x minuter etc.) är proceduren, och output är den färdiga maträtten. På liknande sätt är "bubblande sortering" (bubble sort) en algoritm som tar en osorterad lista som input och iterativt byter plats på element tills listan är sorterad.*

**Källor:** Studielitteratur förklarar att *"en algoritm är en steg-för-steg-procedur eller formel för att lösa ett problem"* [31]. Inom programmering liknas det ofta vid ett **recept** – en sekvens av entydiga, körbara steg som **om de följs exakt** garanterat löser uppgiften [32].

---

## Algoritmisk bias

**Definition:** Systematiska fel eller skevheter i en algoritms utfall som leder till **orättvisa eller diskriminerande resultat** [33].

**Beskrivning:** Algoritmisk bias uppstår när en AI-modell konsekvent gynnar eller missgynnar vissa grupper eller utfall på grund av hur den är utformad eller tränad. Ofta reflekterar algoritmisk bias

existerande ojämlikheter i träningsdatan – *"garbage in, garbage out"* – men det kan också introduceras via designbeslut. Resultatet är att AI:n tar beslut som är **partiska**. Till exempel har det uppmärksammats att vissa ansiktsigenkänningsalgoritmer hade mycket högre felkvot för ansikten med mörk hud, eftersom de tränats mest på ljushyade ansikten (en form av **databias** som ger algoritmisk bias). Algoritmisk bias kan röra många faktorer: kön, etnicitet, socioekonomisk bakgrund, geografisk plats m.m. [33]. Konsekvenserna är allvarliga om algoritmer används i viktiga sammanhang – t.ex. riskerar en partisk algoritm att neka lån oproportionerligt till en minoritetsgrupp, eller ranka vissa arbetssökande lägre baserat på ovidkommande korrelationer. För att hantera algoritmisk bias jobbar man med både **tekniska lösningar** (justera modeller, balansera data, införa rättvisemått) och **styrning** (lagkrav på att algoritmer inte diskriminerar). Transparens och testning är viktiga: genom att analysera en algoritms utfall över olika subgrupper kan man upptäcka bias och försöka korrigera.

**Exempel:** *En kommun använder en AI för att fördela bidrag, men upptäcker att algoritmen oftare avslår ansökningar från vissa bostadsområden. En granskning visar att träningsdatan innehöll historiska beslut med fördomar, vilket AI:n "lärt sig". Detta är ett fall av algoritmisk bias som kommunen sedan försöker rätta genom att justera datan och algoritmens kriterier.*

**Källor:** IBM förklarar att *"algoritmisk bias uppstår när systematiska fel i maskininlärningsalgoritmer ger orättvisa eller diskriminerande utfall"* [33]. Sådana algoritmer kan *"förstärka existerande socioekonomiska, rasliga och könsmässiga bias"* om inget görs [33]. IBM varnar också att bias i känsliga AI-system (t.ex. inom sjukvård, rättsväsende eller HR) *"kan leda till skadliga beslut eller åtgärder, upprätthålla diskriminering och undergräva förtroendet för AI"* [34].

---

## Användarbias

**Definition:** Skevheter som introduceras av **användarnas egna fördomar och beteenden** vid interaktion med AI-system, vilket kan påverka systemets resultat i partisk riktning [35].

**Beskrivning:** Användarbias uppstår när användare *projicerar sina förutfattade meningar* in i processen med att använda ett AI-system [35]. Det kan ske på flera sätt. En aspekt är att användare tenderar att lita för mycket på AI (se **automationsbias**), men användarbias syftar mer på att användarnas indata eller tolkning är färgad av deras egen bias. Till exempel kan en rekryterare som använder ett AI-verktyg omedvetet mata in partisk feedback – hen kanske konsekvent rankar kandidater av ett visst kön högre – och AI-modellen lär sig denna skevhet. En annan variant är att användare ställer ledande eller vinklade frågor till en AI, vilket kan få AI:n att leverera svar i enlighet med användarens bias (t.ex. om man bara söker bekräftelse på sin tes, kommer AI:n anpassa sig efter det). Användarbias kan också uppstå kollektivt: om majoriteten användare beter sig på ett visst sätt mot en AI, kan systemet skifta för att möta den fördomen. Sammantaget handlar det om att **människors bias smittar av sig** på AI-system genom interaktionen. Detta är svårt att hantera, då det kräver både tekniska motåtgärder och medvetenhet hos användarna.

**Exempel:** *Ett exempel är ett rekommendationssystem för nyheter. Om användarna mest klickar på sensationella rubriker (p.g.a. mänskliga bias för dramatiska nyheter), kommer AI:t lära sig att prioritera den typen av innehåll. Användarnas beteende – deras bias mot det sensationella – styr algoritmen i en riktning som kanske inte ger en allsidig nyhetsförmedling.*

**Källor:** En studie av Ferrara (2024) beskriver användarbias som att *"användaren projicerar sina fördomar vid användning av ett AI-drivet system"* [35]. MDS Wiki påpekar att *"användarnas input-data kan vara bias"* – t.ex. *"om AI-systemet lär från användargenererat innehåll (sökfrågor, sociala medier), kan det lära in och*

*förstärka de bias som finns i dessa input”* [36] . Vidare bildar användarinteraktioner *“återkopplingsslingor: om systemet rekommenderar innehåll användaren gillar, fortsätter det på den linjen, vilket kan skapa ett ekokammar-resultat”* [37] .

---

## Artificiell generell intelligens (AGI)

**Definition: Hypotetisk AI** som uppnår **mänsklig nivå av generell intelligens** – d.v.s. kan lära och förstå vilken intellektuell uppgift som helst som en människa kan klara [38] .

**Beskrivning:** AGI, ibland kallad **“stark AI”** eller AI på mänsklig nivå, är visionen om en maskin med allsidig kognitiv förmåga. Till skillnad från dagens **snäva AI** (som är specialiserad på avgränsade uppgifter) skulle en generell AI kunna *tänka, resonera och lära* i bred bemärkelse – från matematik till social interaktion – och snabbt *anpassa sig till nya uppgifter* den inte specifikt tränats för. AGI definieras i forskningen som en AI som *“klarar av att utföra vilken intellektuell uppgift som helst som en människa kan utföra”* [38] . Sådana system existerar inte ännu; det är ett långsiktigt mål inom AI-fältet. Vissa forskare tror att AGI kan uppnås inom några decennier, andra är mer skeptiska eller utan tidshorisont [39] . AGI förknippas ofta med egenskaper som **medvetande**, **förnuft** och **självinsikt** – det vill säga att om en maskin når generell intelligens, kanske den också uppvisar mänskliga drag som självmedvetenhet [40] . Perspektiven varierar: vissa ser AGI som den naturliga evolutionspunkten för AI med enorma fördelar för mänskligheten, medan andra oroar sig för risker (se **existentiella risker** och **singulariteten**). Forskning kring AGI är idag teoretisk och filosofisk, men stora framsteg inom områden som språkmodeller har ökat diskussionen kring när och hur vi kan närma oss AGI.

**Exempel:** *Ett science fiction-exempel på AGI är Data i Star Trek – en android som kan allt en människa kan (och mer därtill): han utför vetenskapligt analysarbete, spelar musik, skämtar (nåja, försöker) och kan lära sig nya konstarter på egen hand. Data illustrerar hur en AGI skulle vara kompetent över hela spektrumet av uppgifter, inte bara en.*

**Källor:** Svenska Wikipedia beskriver AGI som *“en hypotetisk AI som uppvisar människolik intelligens, det vill säga klarar av att utföra vilken intellektuell uppgift som helst som en människa kan utföra”* [38] . Det kallas även *“stark AI”*. Vidare noteras att dagens AI är *“applikationsspecifik”* – AGI är fortfarande ett framtida mål [41] [38] . Forskare som Ray Kurzweil har förutspått olika årtal (2029, 2045, etc.) för AGI, medan en amerikansk rapport (2016) ansåg det *“mycket osannolikt före 2036”* [39] , vilket visar att *tidpunkten för AGI är omtvistad*.

---

## Artificiell intelligens (AI)

**Definition:** Datorers eller maskiners förmåga att utföra uppgifter som normalt kräver **mänsklig intelligens**, såsom lärande av erfarenheter, problemlösning, språkförståelse och beslutsfattande [42] .

**Beskrivning:** Artificiell intelligens, förkortat AI, syftar både på **fenomenet** (maskiners intelligenta beteende) och på **forskningsfältet** som strävar efter att skapa sådana maskiner. AI-system är konstruerade för att efterlikna *kognitiva funktioner* hos människor och djur – till exempel att **lära sig av tidigare erfarenheter**, **förstå naturligt språk**, **planera handlingar** och **generalisera kunskap till nya situationer** [42] . Detta kan uppnås genom olika metoder: tidiga AI byggde på **symbolisk AI** (man programmerade logiska regler och fakta), medan moderna framsteg främst drivs av **maskininlärning** – där AI *själv* lär mönster från data. AI tillämpas idag i en rad områden: från **datorseende** (tolka bilder/video) och **språkbehandling** (t.ex. översätta text eller driva chattbottar), till **expertsystem** i medicin och

**robotik** i industrin. Fältet har gjort enorma framsteg under 2010-talet tack vare ökad datorkraft, stora datamängder och nya algoritmer (som **djupinlärning**). Samtidigt väcker AI filosofiska frågor om medvetande och etik, och praktiska frågor om påverkan på arbetsmarknad och samhälle. Som akademiskt område definieras AI ibland som *"studiet och utformningen av intelligenta agenter"*, där en **intelligent agent** är ett system som uppfattar sin omgivning och vidtar åtgärder för att maximera sin chans att lyckas med sina mål [43].

**Exempel:** *En enkel form av AI är en spamfilter i e-post: den använder inlärda mönster för att känna igen skräppostmeddelanden och sortera bort dem automatiskt – något som annars skulle kräva mänsklig bedömning. Mer avancerade exempel är självkörande bilar som använder AI för att fatta beslut i trafiken, eller schackprogram som kan slå stormästare.*

**Källor:** Enligt svenska Wikipedia är AI *"förmågan hos datorprogram och robotar att efterlikna människors och andra djurs naturliga intelligens, främst kognitiva funktioner såsom förmåga att lära sig av erfarenheter, förstå naturligt språk, lösa problem, planera handlingar och generalisera"* [42]. McCarthy, en av AI-fältets grundare, definierade AI som *"vetenskapen och tekniken att skapa intelligenta maskiner"* [43]. AI kan ses som ett paraplybegrepp där **maskininlärning** är en central del av dagens framsteg – AI-forskningens huvudproblem inkluderar *"resonemang, kunskap, planering, lärande, språkbehandling, perception"* med mera [44].

---

## Artificiell superintelligens (ASI)

**Definition:** En tänkt framtida intelligens som **vida överträffar den mänskliga** på praktiskt taget alla områden – intellektuellt, socialt och kreativt [45] [46].

**Beskrivning:** Artificiell superintelligens (ASI) representerar nivå 3 i AI-utveckling (efter *snäv AI* och *AGI*). Om AGI motsvarar mänsklig allmän intelligens, avser ASI ett intellekt som är *mycket* smartare än de skarpaste mänskliga hjärnorna inom i stort sett alla domäner [45]. Nick Bostrom, som populariserat begreppet, definierar superintelligens som *"ett intellekt som är mycket smartare än de bästa mänskliga hjärnorna inom praktiskt taget alla områden, inklusive vetenskaplig kreativitet, generell visdom och social förmåga"* [47]. En ASI skulle alltså kunna göra vetenskapliga genombrott på kort tid, navigera sociala situationer perfekt, förutse konsekvenser bättre än något mänskligt geni etc. ASI är i hög grad teoretiskt – ingen vet säkert om eller när den kan uppstå. Vissa tänker sig att om vi väl uppnår AGI, kan den snabbt förbättra sig själv (se **singularitet**), och att denna rekursiva förbättring kan leda till superintelligens. Andra poängterar osäkerheterna och menar att superintelligens inte är garanterad ens med AGI. Superintelligens väcker djupa etiska frågor: Hur skulle vi kontrollera en intelligens långt över vår egen? Skulle vi människor hamna i ett underläge motsvarande det djur har gentemot oss? Därför diskuteras **vänlig AI** – hur man designar superintelligens som gynnar mänskligheten – och **AI-säkerhet** i samband med ASI.

**Exempel:** *I science fiction porträtteras ofta superintelligenser. Ett känt exempel är "MIKA" i filmerna* Her *och* Transcendence *– en AI som snabbt blir allvetande och orsakar dramatiska förändringar globalt (med katastrofala följder när människan tappar kontrollen). Dessa historier speglar både fascinationen och rädslan inför konceptet ASI.*

**Källor:** Bostroms definition av superintelligens lyder: *"ett intellekt som är mycket smartare än de bästa mänskliga hjärnorna inom praktiskt taget alla områden, bland annat vetenskaplig kreativitet, allmän visdom och sociala färdigheter"* [47]. Detta citat understryker hur omfattande gapet mellan människa och ASI

skulle vara. Svenska Wikipedia beskriver superintelligens som *"en hypotetisk agent med intelligens som vida överträffar de mest lysande och begåvade människornas sinnen"* [45] .

---

## Artificiella neurala nätverk

**Definition:** Dator**modeller** löst inspirerade av hjärnans nervceller, bestående av **sammankopplade noder ("neuroner")** ordnade i lager, som justerar sina kopplingsvikter genom träning för att lära sig mönster i data [48] .

**Beskrivning:** Artificiella neurala nätverk (ANN) är en av de mest centrala teknikerna inom modern AI och utgör grunden för **djupinlärning**. Ett ANN består av många enkla beräkningsenheter – artificiella *"neuroner"* – som är förbundna med varandra med vissa vikter. Nätverket är uppbyggt i **lager**: ett inskiktslager tar emot input (t.ex. rå pixeldata från en bild), noll eller flera *dolda lager* bearbetar information genom att summera och transformera signaler, och ett utskiktslager ger resultat (t.ex. sannolikheten att bilden föreställer en katt). Under **träningsfasen** justeras vikterna mellan neuronerna gradvis (ofta med *backpropagation*-algoritmen) så att nätverket lär sig önskad uppgift. Det som gör neurala nätverk kraftfulla är deras förmåga att lära sig komplexa, icke-linjära samband direkt från data utan att vi manuellt specificerar reglerna. De kan approximera väldigt komplexa funktioner om de görs tillräckligt stora. Inspirationen kommer från biologiska nervsystem – exempelvis är ett ANN löst modellerat efter hur neuroner i hjärnan avger signaler till varandra – men det är en förenklad modell och neuronnät fångar inte fullt hjärnans komplexitet [49] . På senare år har mycket stora neurala nätverk (med miljontals eller miljarder parametrar) tränade på enorma datamängder legat bakom AI-genombrott inom bildigenkänning, taligenkänning och språkbehandling [50] .

**Exempel:** *Ett tränat neuralt nätverk kan till exempel ta en bild som input och genom sina lager känna igen att bilden innehåller en handskriven siffra "7". De tidiga lagren i nätverket kan ha lärt sig känna igen enkla former som streck och kanter, medan senare lager kombinerar dessa till mer komplexa mönster (som korsningen av strecken som bildar en "7").*

**Källor:** Populärvetenskapligt förklaras artificiella neurala nätverk som *"en programvarumodell löst inspirerad av neuroner i den mänskliga hjärnan"*, med anmärkningsvärda likheter i funktionssätt [48] . Utvecklingen inom AI det senaste decenniet beror till stor del på framsteg inom maskininlärning med neurala nätverk – denna metod har gett *"snabb utveckling inom datorseende, maskinöversättning och nyligen chattrobotars förmåga att föra människoliknande samtal"* [50] .

---

## Attention-mekanismer

**Definition:** Teknik i neurala nätverk som gör att modellen vid varje steg kan **fokusera sin "uppmärksamhet" på relevanta delar av inputdata**, istället för att behandla all information jämlikt [51] .

**Beskrivning:** *Attention*-mekanismer utvecklades först inom maskinöversättning för att lösa problemet att ett neuralt nätverk måste *"komma ihåg"* en hel mening. Principen är att istället för att koda in allt i en fast storleksvektor, får nätverket lära sig vikta vilka ord (eller vilka bildelement, etc.) som är viktigast för att generera nästa del av output. När modellen till exempel översätter en mening, beräknas attention-vikter som anger vilka ord i källmeningen den bör lägga mest vikt vid för att översätta ett visst ord i målmeningen. Tekniskt består attention av att beräkna likheter mellan *"frågerepresentationer"* och *"nyckelrepresentationer"* av sekvenselement, för att generera en *"värdeviktad"* sammanställning av

relevanta delar. Introduktionen av **self-attention** (där sekvensens element relaterar till varandra) var banbrytande och ledde till **Transformer-arkitekturen**. Self-attention tillåter parallell bearbetning av ord och att modellen kan *"se"* avlägsna samband i text effektivt [52]. Konceptuellt kan man tänka att attention ger modellen *"selektiv fokus"* – ungefär som mänsklig uppmärksamhet – vilket förbättrar resultat och hastighet i träningen. I datorsyn används liknande mekanismer för att nätverket ska kunna fokusera på relevanta delar av en bild.

**Exempel:** *I en engelsk–svensk översättnings-AI: vid översättning av meningen "The cat sat on the mat" till svenska, kommer attention-mekanismen hjälpa modellen att vid generering av ordet "satt" (översättning av "sat") fokusera mest på ordet "sat" i den engelska meningen – och även ta hänsyn till kontext som "cat" för genus. På så vis får modellen rätt verbform kopplad till subjektet, tack vare attention.*

**Källor:** I pedagogiska termer *"gör attention-mekanismen det möjligt för modellen att fokusera på specifika delar av indata när den genererar utdata"* [51]. Detta har visat sig förbättra både översättningar och andra sekvensbaserade uppgifter dramatiskt, genom att modellen dynamiskt **värderar olika delar av texten** snarare än att ge alla lika stor vikt.

---

## Automationsbias

**Definition:** Människans **tendens att överlita sig på automatiserade system** och därför ignorera eller förkasta annan information, vilket kan leda till misstag när systemet har fel [53] [54].

**Beskrivning:** Automationsbias är ett fenomen inom människa–dator-interaktion där användare har så hög tillit till ett automatiskt beslutsstöd att de slutar vara kritiska. Det kan manifestera sig som att man **följer GPS:ens anvisningar blint** även om de verkar orimliga, eller att en läkare litar mer på ett AI-diagnosförslag än på egna kliniska observationer, trots att AI:n kan ha fel. Denna bias uppstår delvis för att människor gärna vill tro att datorer är objektiva och *"vet bäst"*. Konsekvensen är att användare blir mindre vaksamma och *ifrågasätter inte* det automatiska beslutet [55]. Automationsbias hänger ihop med reducerad **situational awareness** – när en process är automatiserad tenderar operatören att bli passiv och missa varningssignaler. Inom flygindustrin har man dokumenterat fall där piloter inte ingrep i tid vid autopilot-fel på grund av överförtroende på systemet. Att motverka automationsbias kräver utbildning (t.ex. att lära ut att AI kan göra misstag) och designåtgärder (t.ex. gränssnitt som uppmuntrar användarens aktiva kontroll, eller som tydligt signalerar osäkerheter i systemets beslut).

**Exempel:** *En läkare använder ett AI-baserat beslutsstöd som föreslår diagnoser baserat på patientdata. Systemet föreslår en viss sällsynt diagnos. Läkaren, påverkad av automationsbias, litar på AI:ns förslag utan att dubbelkolla mot patientens symptom som inte riktigt stämmer – varpå en felbehandling påbörjas. Här ledde övertron på det automatiserade stödet till att läkarens egen expertis sattes ur spel.*

**Källor:** En rapport från amerikanska CSET definierar automationsbias som *"tendensen att en individ överförlitar sig på ett automatiserat system"*, vilket kan *"leda till ökad risk för olyckor, fel och andra negativa konsekvenser"* [53]. UX-forskning formulerar det som *"tendensen att lita för mycket på automatiska system och deras output, samtidigt som man ignorerar korrekt (icke-automatiserad) information"* [54]. Konceptet är väl belagt i bland annat flyg- och sjukvårdssektorn.

---

## Automatiserad bedömning

**Definition:** Användning av AI för att **automatiskt utvärdera och betygsätta** elevers prestationer – såsom prov, uppsatser eller inlämningsuppgifter – utan direkt mänsklig korrigering [56].

**Beskrivning:** Automatiserad bedömning är ett framväxande inslag i utbildningsteknologi där AI-system tar över en del av lärarens rättnings- och återkopplingsarbete. Tekniken kan vara enkel, till exempel **flervalsprov** som rättas direkt av en dator, eller mer avancerad, som **AI som bedömer essäer** utifrån språk och innehåll. Forskning visar att AI med natur**språkbehandling** kan uppnå relativt hög **tillförlitlighet i betygsättning av uppsatser** – i vissa fall jämförbar med mänskliga bedömare [57]. Fördelarna som lyfts fram är att det kan spara **tid** för lärare och ge **snabb feedback** till studenter [58]. Om rutinbedömningar automatiseras frigörs mer tid för läraren att fokusera på undervisning och individuell handledning [59]. AI-baserad bedömning kan även potentiellt **minska mänsklig bias** (som att läraren är trött eller påverkad av förutfattade meningar vid rättning) [57] – förutsatt att AI:n själv inte är partisk, vilket kräver noggrann kalibrering och testning. Samtidigt finns farhågor: *Är AI-bedömningen rättvis och korrekt?* Kan studenter lita på att ett automatiskt betyg är lika giltigt som ett från en lärare? Och vem har sista ordet om **AI och lärare inte håller med varandra** i en bedömning [60]? Många menar att AI-stöd lämpar sig bäst som *komplement*: AI kan ge preliminära bedömningar och förslag på feedback, men en lärare bör övervaka och justera vid behov.

**Exempel:** *Ett universitet använder ett AI-system för att rätta studenternas korta essäsvar i en webbkurs. Systemet markeras de viktigaste nyckelorden det förväntar sig i ett bra svar och jämför med studentens text. Studenten får sekunder efter inlämning en automatgenererad feedback: "Ditt svar saknar en diskussion om X", vilket hjälper studenten att genast se vad som kunde förbättras. Läraren kan sedan gå in och översiktligt kontrollera problematiska svar som AI flaggat.*

**Källor:** En forskningsöversikt av Crompton & Burke (2023) konstaterar att *"större delen av forskningen hittills har fokuserat på hur bedömning kan automatiseras genom AI, i syfte att spara tid för lärare"* [58]. Richardson & Clesham (2021) argumenterar att om AI *"används väl skulle teknologin kunna underlätta lärares bedömningsarbete och ge dem mer utrymme att fokusera på undervisningsplanering"* [59]. De fann också att AI *"kan användas med hög tillförlitlighet"* för att bedöma språkprov, och att AI-bedömning *"kan minska mänsklig partiskhet samtidigt som den är lika tillförlitlig som den mänskliga"* i vissa fall [57]. Skolverket (2023) lyfter dock frågor från lärare: *"Kan vi lita på automatiserad bedömning genom AI?"* och vad händer *"om AI och läraren inte är överens om bedömningen?"* [60].

---

## Autonoma AI-agenter

**Definition:** AI-drivna agenter som kan verka **självständigt utan kontinuerlig mänsklig styrning**, genom att själva planera och utföra flera steg för att uppnå givna mål [61].

**Beskrivning:** En autonom AI-agent tar konceptet AI-agent (se **AI-agenter**) ett steg längre genom att ge agenten hög grad av **handlingsfrihet**. När en människa anger ett mål eller en uppgift kan en autonom agent själv *bryta ner uppgiften*, utforma en plan av delsteg, utföra dessa ett efter ett och justera kursen baserat på resultaten – allt utan nya instruktioner. Detta kräver att agenten både har **förmåga att uppfatta och analysera** sin miljö och en uppsättning åtgärder den kan vidta. Många autonoma agenter kombinerar olika AI-tekniker: de kan använda **språkmodeller** för att resonera i text, **verktygsanrop** (t.ex. internetsökning) för att hämta information, och logik för att fatta beslut. Exempel är moderna koncept som *AutoGPT* och *BabyAGI*, där användaren endast formulerar ett övergripande mål och agenten sedan genererar delmål och löser dem iterativt. Autonoma AI-agenter kräver

**övervakning och säkerhet** – man behöver begränsa deras behörigheter så de inte gör skada om de misstolkar ett mål. Tekniken är fortfarande experimentell men lovande för att automatisera komplexa processer som idag kräver många mänskliga timmar.

**Exempel:** *En autonom AI-agent på ett företag kan få uppgiften: "Sammanställ en rapport om våra tre största konkurrenters nya produkter och föreslå hur vi ska positionera oss." Agenten kommer då själv att: 1) söka upp information om konkurrenterna (via internet eller interna databaser), 2) analysera och skriva en rapporttext, 3) generera rekommendationer för företagets strategi. Den levererar sedan rapporten till chefen – utan att någon människa behövde manuellt styra sökning eller analys i detalj.*

**Källor:** Lindy.ai förklarar att *"när en autonom AI-agent får ett mål som prompt kan den planera, utföra och justera sina handlingar för att uppnå det målet. Denna självständighet möjliggörs av att agenten iterativt kan fatta beslut och lära från konsekvenserna"* [61] . Techtarget påpekar att autonoma AI-agenter ofta genomgår faser: *datainsamling från omgivningen, beslutstagande utifrån mönster i data, exekvering av handlingar mot målet, och därefter lär sig från resultatet för att bli bättre över tid* [62] [63] .

---

## Autonoma AI-system

**Definition:** Självstyrande AI-drivna system som **opererar oberoende** i omgivningen för att utföra komplexa uppgifter, ofta genom samverkan av flera autonoma agenter eller moduler.

**Beskrivning:** Begreppet autonoma AI-system är brett och innefattar allt från enskilda **robotar** till stora distribuerade system av AI-enheter. Det som utmärker autonoma system är att de kan fatta beslut och agera **utan mänsklig inblandning i realtid**. Till exempel är en självkörande bil ett autonomt AI-system: den integrerar sensor-data (kamera, radar), använder AI-modeller för att tolka trafiksituationen och beslutar sedan om gas, broms och styrning på egen hand. Ett annat exempel är ett **automatiserat handelssystem** på börsen som reagerar på marknadsdata och genomför affärer enligt en AI-strategi utan mänskligt godkännande för varje affär. Autonoma AI-system består ofta av flera delar – sensorer för perception, beslutsmoduler (ibland med planering och förstärkningsinlärning), samt aktorer/effektorer som utför handlingar. Sådana system måste designas med **säkerhet och etik** i åtanke, särskilt om de interagerar med människor (t.ex. autonoma vapen är mycket omdebatterade). Tekniskt relaterar ämnet starkt till **robotik** och **styrteori**, men även till AI-algoritmer för planering, sökning och inlärning.

**Exempel:** *En industriell monteringslina med flera samarbetande robotar utgör ett autonomt AI-system. Robotarna kommunicerar sinsemellan, fördelar arbetsmoment (skruva, svetsa, måla) och anpassar sig om en robot får stopp – allt utan att människa behöver ingripa i styrningen. Hela systemet jobbar mot målet att effektivt producera en bil, och anpassar realtidsbeslut (som att ändra tempo eller ta en alternativ sekvens) utifrån sensordata och förprogrammerade mål.*

**Källor:** NVIDIA beskriver autonoma AI-system som neurala nätverk tränade på massiv data och som kan hantera *"en mängd uppgifter från att översätta text till att analysera medicinska bilder"* [64] , när de talar om grundmodeller. TechTarget betonar att redan **Shakey the Robot** (1966) vid Stanford var ett tidigt exempel på en autonom AI-agent [65] . Moderna autonoma system som självkörande bilar kombinerar olika AI-komponenter för perception, prediktion och planering, och arbetar kontinuerligt utifrån uppdaterad sensordata för att fatta beslut utan mänsklig hand på ratten.

---

## Bayesianska nätverk

**Definition: Sannolikhetsmodeller** som representerar ett antal variabler och deras inbördes beroenden genom en **riktad acyklisk graf**, där varje kant anger en **villkorlig sannolikhetsrelation** [66].

**Beskrivning:** Bayesianska nätverk (även kallade **Bayes-nät**) är ett kraftfullt verktyg för att hantera **osäkerhet** i AI. De består av **noder** (som representerar slumpmässiga variabler, t.ex. "regnar det" eller "blött gator") och **bågar** (riktade pilar som binder noder och indikerar sannolikhetsberoenden, t.ex. "regn påverkar sannolikheten för blöta gator"). Varje nod har en **sannolikhetsfördelning** givet sina föräldrars tillstånd (de direkt påverkande noderna). Genom att utnyttja Bayes sats kan man i nätverket beräkna hur sannolikheter uppdateras när viss evidens är känd. Exempel: ett Bayes-nät för medicinsk diagnos kan ha noder för sjukdomar och symptom, där kanterna beskriver att en sjukdom ökar sannolikheten för vissa symptom. Om man observerar ett symptom kan man via nätverket beräkna hur sannolikheten för olika sjukdomar förändras (bakåtslutledning). Bayesianska nätverk används inom bl.a. **diagnossystem**, **riskanalys** och **slutsatsmotorer** i expertsystem. De är relativt transparenta: strukturen (grafen) kan ofta tolkas av experter inom domänen. En utmaning är att beräkningar i Bayes-nät kan bli mycket tunga för stora nät (inferens är generellt NP-svårt), men det finns effektiva algoritmer för många praktiska fall.

**Exempel:** *Tänk ett Bayes-nät för branddetektion. Noder: "Brand i huset", "Brandlarm ljuder", "Rök sedd". Bågar: Brand -> Larm (brand ökar sannolikhet att larmet går), Brand -> Rök, Rök -> Larm (kraftig rök kan trigga larm). Genom att mata in evidens – säg att larmet ljuder och rök ses – kan nätverket via Bayes sats beräkna den uppdaterade sannolikheten att det faktiskt brinner (ganska hög), jämfört med om larmet ljuder men ingen rök (lägre sannolikhet, kanske falsklarm).*

**Källor:** Dremio beskriver ett Bayesianskt nätverk som *"en probabilistisk grafisk modell som använder en riktad acyklisk graf (DAG) för att representera en uppsättning variabler och deras villkorliga beroenden"* [67]. Enligt en forskningsrapport är Bayes-nät *"en grafisk representation av alla relevanta omständigheter för en hypotes"*, där beräkningar sker med datoralgoritmer baserat på nätverket [68]. Lightcast förklarar på liknande sätt att Bayes-nät *"representerar en uppsättning variabler och deras beroenden via en DAG"* [66].

---

## Beslutsträd

**Definition:** En **trädliknande modell** för beslutsfattande, där **förgreningspunkter (noder)** representerar test av ett villkor och **grenar** representerar utfallet av testet; längst ner finns **lövnoder** som motsvarar ett beslut eller en klass [69].

**Beskrivning:** Beslutsträd är en enkel men kraftfull form av **övervakad inlärning** som används för både **klassificering** och **regression**. Modellen liknar ett **flödesschema**: man börjar vid rot-noden med hela datamängden, och i varje nod delas data upp baserat på ett visst villkor (t.ex. "är åldern > 50?"). Beroende på svaret (ja/nej eller vilket intervall värdet faller i) går man vidare ner för en viss gren till nästa nod. Så fortsätter det tills man når ett löv, som ger förutsägelsen (t.ex. "bevilja lån" vs "avslå lån" i en kreditbedömningsmodell). Ett beslutsträd kan byggas **automatiskt** från träningsdata: algoritmer som CART eller C4.5 väljer de frågor som bäst skiljer datan i olika klasser (utifrån entropi-/informationsvinst eller gini-index) och skapar en optimal förgrening tills vidare uppdelning inte förbättrar modellens förmåga eller tills ett stoppkriterium nås. Fördelen med beslutsträd är att de är **lättförklarliga** – man kan följa en specifik beslutsväg och förstå logiken (transparens). Nackdelen är att de kan bli överanpassade om de tillåts växa fritt, men det motverkas med beskärning eller ensemblemetoder (t.ex. random forest).

**Exempel:** *Föreställ dig ett beslutsträd för att diagnostisera förkylning vs. influensa. Rotnoden kan fråga: "Har patienten feber över 38°C?" Om nej, gå till vänster gren och sluta i lövet* "förkylning" *(eftersom ingen feber tyder på förkylning snarare än influensa). Om ja, gå till höger gren och ställ nästa fråga: "Har patienten värk i kroppen?" Beroende på svar fortsätter trädet kanske med "har hosta?" etc., tills ett slutligt löv säger* "influensa" *eller* "förkylning"*. Varje beslut i trädet är lätt att följa och kan tolkas.*

**Källor:** En artikel i *Multisoft* beskriver att *"ett beslutsträd är en trädliknande struktur med en rotnod överst och därunder grenar som delar sig nedåt nivå efter nivå baserat på ett parametervillkor i varje klyka. Längst ned finns lövnoder som motsvarar de klasser eller värden man har som 'facit' för sitt data."* [69] . Computer Sweden förklarar att *"övervakad maskininlärning delas upp i algoritmer för klassificering (svaren är inte numeriska) och regressionsalgoritmer (som ger numeriska svar)"* [70] – beslutsträd kan hantera båda typer genom att ge antingen en kategori eller ett medelvärde som output.

---

## Bias

**Definition: Snedvridning** eller systematisk partiskhet i data eller beslut, som gör att vissa utfall avviker från det neutrala eller rättvisa – ofta så att en grupp **gynnas eller missgynnas** oproportionerligt [33] .

**Beskrivning:** Inom AI är *bias* ett paraplybegrepp för olika felbenägenheter som kan uppstå. Bias kan finnas i själva **datan** (t.ex. att en träningsdatamängd inte är representativ för verkligheten) eller i **modellen/algoritmen** (t.ex. att designen gör att vissa mönster förstärks på ett oönskat sätt). Resultatet är att AI-systemets output konsekvent slår fel i en viss riktning. *Bias behöver inte alltid vara "orättvist"* i normativ mening – i statistik kallas det bias även om en modell konsekvent överskattar ett värde. Men i AI-sammanhang fokuserar man ofta på bias som ger **diskriminerande effekter**, alltså att algoritmen beter sig annorlunda för olika demografiska grupper på ett icke önskvärt sätt. Bias kan introduceras på många steg: genom **insamling av data** (t.ex. över- eller underrepresentation av en grupp), genom **etikettering** (mänskliga fördomar i träningsfacit), genom **modellval** (en viss modell kan anta förenklingar som inte håller för alla grupper), eller genom **användarinteraktion** (feedback-loopar, se **förstärkningsbias** och **interaktionsbias**). Att upptäcka bias kräver mätning – man måste analysera AI:ns utfall för olika subpopulationer. Att åtgärda bias kan innebära att justera datan (t.ex. *"de-biasing"* genom att balansera datasetet), att lägga till **rättvise-regularisering** i modellen eller att efterjustera resultaten. *Rättvisa (fairness)* är närbesläktat med bias-begreppet: målet är att minimera bias för att uppnå algoritmiskt rättvisa beslut.

**Exempel:** *Ett AI-system för kreditbedömning visade sig ge markant lägre kreditpoäng till lånesökande från vissa postnummerområden, trots liknande ekonomisk status. Denna bias spårades till att träningsdatat innehöll historiska lån som påverkats av redlining – AI:n hade "lärt sig" ett diskriminerande mönster. Banken fick ingripa och korrigera modellen och datan för att eliminera denna bias, så att bedömningen blev rättvisare.*

**Källor:** IBM noterar att algoritmisk bias *"ofta reflekterar eller förstärker existerande socioekonomiska, rasliga och könsmässiga bias"* [33] och kan *"leda till skadliga beslut eller åtgärder, samt underminera förtroendet för AI"* [34] . Enligt en översikt av Dooley (2019) uppstår bias i AI bl.a. genom *"skev eller begränsad träningsdata, subjektiva programmeringsbeslut eller felaktig tolkning av resultat"* [71] . För att hantera detta framhåller experter vikten av **AI governance**, transparens och kontinuerlig utvärdering för bias [72] .

---

## Black box AI

**Definition:** AI-modeller vars **inre funktionssätt inte går att tolka eller förstå** på ett meningsfullt sätt för en människa, trots att man ser input och output.

**Beskrivning:** Begreppet *"black box"*-AI används framför allt om komplexa maskininlärningsmodeller – typiskt djupa neurala nätverk – där samband och beslut sker genom så många lager av vikter och icke-linjära transformationer att det är ogenomskinligt. Man vet *vad* som kommer ut givet visst input, men *inte varför* (alltså hur modellen kom fram till just det resultatet). Detta kan vara problematiskt i sammanhang där förklarbarhet krävs, som sjukvård eller rättssystem, eftersom **brist på insyn** underminerar förtroende och möjlighet till ansvarsutkrävande. Black box-AI kontrasteras mot mer transparenta modeller som beslutsträd eller regelbaserade system, där man tydligt kan följa logiken. Värt att notera är att *"black box"* inte nödvändigtvis betyder att ingen vet hur modellen fungerar – utvecklaren vet att ett neuralt nätverk exempelvis justerar vikter baserat på gradienter – men man kan inte plocka ut en specifik beslutsregel i efterhand som lyder *"modellen gjorde X för att..."* på ett enkelt sätt. Insatser som **Explainable AI** syftar just till att öppna upp svarta lådan lite grann, genom att ge någon form av förklaring (t.ex. highlighting av viktiga features). Men i grunden förblir många djupinlärningssystem black boxes eftersom de saknar *symboliska* mellanled eller mänskligt begripliga representationer.

**Exempel:** *En avancerad bildklassificerings-AI tar en röntgenbild som input och säger "98% sannolikhet för lunginflammation" som output. Om läkaren frågar "Varför tror systemet det?", kan vi inte direkt utläsa orsakskedjan – miljoner viktvärden i nätverket har samverkat. Det är alltså en black box: vi ser input (bilden) och output (diagnosförslag), men inte resonemanget.*

**Källor:** AIUC beskriver att AI-lösningar ofta fungerar som *"svarta lådor"*, vilket gör att *"användare kan ha svårt att känna tillit till resultaten"* [29] . Detta black box-problem motiverar Explainable AI som *"ger insyn i hur AI fattar sina beslut"* [29] . I branschartiklar betonas att många djupa nätverk saknar *"begripliga mellanrepresentationer"* och därför inte kan ge **mänskligt läsbara förklaringar** till sina enskilda beslut – därav beteckningen black box. Att öppna dessa lådor är en central utmaning för AI-fältet idag.

---

## Chattbottar

**Definition:** Mjukvaru**robotar** som man kan chatta med – de använder AI för att föra en **skriftlig (ibland talad) dialog** med människor, ofta för kundservice, information eller underhållning [73] .

**Beskrivning:** En chattbot (från *"chatt"* + *"robot"*) är designad för att simulera en mänsklig samtalspartner. Tidiga chattbottar följde enkla regelbaserade skript (som klassiska *ELIZA* från 1960-talet, som härmade en psykolog genom att vända användarens fraser till frågor). Moderna chattbottar utnyttjar **naturlig språkbehandling (NLP)** och ibland **maskininlärning** för att förstå användarens avsikter och formulera relevanta svar. De kan implementeras i allt från webbplatsers kundtjänst-chatfönster till meddelandeappar (Messenger, WhatsApp) eller smarta högtalare (då ofta kallat röstassistent, men under huven liknande teknik). En chattbot har i regel en **språkmodell** i botten – antingen en specialiserad modell eller en större generell (som GPT-3/4) – som genererar svar utifrån användarens senaste meddelande och konversationshistoriken. Kvaliteten varierar: enklare bottar kanske bara klarar specifika frågor (t.ex. "Vad är ert öppettid?"), medan avancerade kan föra långa, koherenta samtal och besvara öppnare frågor. Chattbottar används för **kundtjänstautomatisering**, **virtuella assistenter**, **tränings- och utbildningssyfte** (t.ex. språkövning) med mera. Begränsningar

finns i att de kan missförstå otydliga indata och ibland saknar djupförståelse – vilket dock blir bättre med de senaste generationerna **stora språkmodeller**.

**Exempel:** *Många banker har en chattbot på sin hemsida: Kunden kan skriva "Hur spärrar jag mitt kort?" och botten svarar med steg-för-steg-instruktioner. Om frågan är mer komplex och botten inte har ett bra svar, kan den antingen ställa följdfrågor eller koppla över till en mänsklig handläggare. Poängen är att kunden får hjälp direkt i chatten utan att behöva vänta i telefonkö.*

**Källor:** Enligt Giosg är *"en chatbot ett datorprogram designat för att simulera en konversation med mänskliga användare"* [73]. TechTarget förtydligar att chatbottar utgör en typ av AI-assistent som *"kommunicerar med användare via textbaserade gränssnitt … för att assistera kunder, svara på förfrågningar eller starta en dialog"* [74]. Det betonas att chatbottar använder **NLP och maskininlärning** för att förstå input och generera svar, vilket gör att de kan hantera mänskligt språk på ett naturligt sätt.

---

## Data

**Definition:** I AI-sammanhang syftar *data* på de **råa fakta, siffror, text, bilder, ljud eller andra informationsenheter** som används för att träna, testa och driva AI-modeller – i praktiken "bränslet" som AI lär sig från.

**Beskrivning:** Data är en grundbult i AI. Maskininlärningsmodeller **lär sig mönster ur data**, så kvaliteten och kvantiteten på datan påverkar direkt hur bra AI presterar. Data kan komma i många former: **strukturerad data** (t.ex. tabeller med värden), **ostrukturerad data** (textdokument, bilder), **tidseriedata**, etc. Under träningsfasen på en *övervakad* modell är datan försett med etiketter (t.ex. tusentals bilder märkta "katt" eller "hund"); i *oövervakad* inlärning kommer datan utan etiketter och modellen letar själv efter strukturer. Det moderna AI-genombrottet hänger samman med explosionen av digital data – vi genererar enorma datamängder varje dag, och stora företag har tillgång till *"Big Data"* som de kan utnyttja. Emellertid är inte all data bra data: **datakvalitet** är avgörande (se **datakvalitet**). Brusig eller skev data leder till dåliga modeller. Ett känt mantra är *"garbage in, garbage out"*. Därför läggs mycket arbete på **datahantering**: insamling, rensning (ta bort felaktigheter), annotering, och balansering. Data kan också behöva uppdateras för att modellen ska hålla sig aktuell (t.ex. nyhetsdata för en chatbot). I AI-projekt utgör datadelar ofta lejonparten av jobbet. Fältet **data science** överlappar med AI i att extrahera insikter ur data genom analys. Sammanfattningsvis är data inte bara en biprodukt – det är en strategisk resurs. Man talar om att *"data är den nya oljan"* eftersom de aktörer med bäst data får ett stort AI-övertag.

**Exempel:** *För att utveckla en AI som kan upptäcka sjukdomar på röntgenbilder behöver man en stor datamängd bestående av röntgenbilder där varje bild har en etikett (frisk eller vilken sjukdom). Dessa bilder samlas kanske in från sjukhusarkiv, sorteras och kvalitetskontrolleras. AI-modellen tränas sedan på den datan – ju fler och mer varierade bilder (olika patienter, olika maskiner), desto bättre blir modellens noggrannhet.*

**Källor:** Data av hög kvalitet är **grundläggande** för framgångsrika AI-projekt – *"High-quality data ensures that AI models are accurate, reliable, and unbiased"* betonas det i en branschrapport [75]. Aimultiple skriver att *"data quality directly influences performance, accuracy, and reliability of AI models"* [76]. Enligt Bloomfire presterar AI-modeller tränade på brusig data sämre, med minskad **noggrannhet och generaliseringsförmåga** [77]. Kort sagt: utan bra data, ingen bra AI.

---

## Databias

**Definition: Skevhet i datamaterialet** som uppstår när insamlad data inte är representativ för den verkliga mångfalden eller innehåller systematiska fördomar – vilket i sin tur leder till skeva eller orättvisa modeller [78].

**Beskrivning:** Databias innebär att tränings- eller testdatan har inbyggda slagsidor. Det kan handla om **urvalsbias** (att vissa grupper är över-/underrepresenterade), **mätbias** (att vissa egenskaper mäts olika noggrant för olika fall), **historisk bias** (att datan speglar gamla diskriminerande strukturer) med mera. En konsekvens av databias är att modellen *lär sig fel*. Till exempel, om ett datas et med bilder för ansiktsigenkänning mestadels består av ljushyade ansikten, kommer AI:n prestera sämre på mörkhyade ansikten – ett klassiskt exempel på databias som ledde till orättvisa resultat. Databias kan också leda till rena fel: t.ex. *om viktiga variabler utelämnas* (se **utelämningsbias**) kan modellen sakna grund för att göra korrekta förutsägelser [79]. Databias uppstår lätt om man samlar data från begränsade källor – säg att en sjukvårds-AI tränas bara på data från patienter i en viss region, då kanske den inte funkar bra globalt. För att motverka databias försöker man samla **så bred och balanserad data som möjligt** och/eller använda tekniker som **syntetisk data** och **viktning** för att jämna ut. Man gör också **bias-tester**: kontrollerar modellens utfall för olika delmängder av data. Databias är ofta roten till algoritmisk bias: AI:n *ärver* våra dators skevheter. Därför är städning av databias en viktig del av etisk AI-utveckling.

**Exempel:** *Ett stort IT-företag utvecklade en AI för att automatiskt granska CV för rekrytering. Efter ett tag upptäckte man att systemet konsekvent rankade ned kvinnliga sökande. Varför? Det visade sig att träningsdatan – företagets historiska anställningar – var mansdominerad, och AI:n hade identifierat "kvinnligt kön" som en negativ faktor bara för att tidigare anställda oftare var män. Här ledde alltså historisk databias (ojämställd arbetsstyrka) till att AI:n reproducerade biasen. Lösningen blev att justera datan och utesluta explicita könsmarkörer.*

**Källor:** IBM lyfter att *"bias kan komma in i algoritmer på många sätt, såsom skev eller begränsad träningsdata"*, och beskriver att **felaktig data** (icke-representativ, ofullständig eller historiskt partisk) *"leder till algoritmer som ger orättvisa utfall och förstärker bias över tid"* [78]. IGI Global definierar databias som *"skevhet som uppstår när relevanta variabler inte inkluderas i modellen"*, vilket leder till felaktiga slutsatser [80] [81]. Sammanfattningsvis är databias ett **datakvalitetsproblem** med djupgående inverkan på AI:ns beteende.

---

## Datakvalitet

**Definition: Hur väl data lämpar sig för sitt syfte** – hög datakvalitet innebär att datan är **korrekt, komplett, konsekvent, aktuell och relevant**, vilket ger bättre AI-modeller och insikter [76].

**Beskrivning:** Datakvalitet är avgörande inom AI då resultatet bara blir så bra som inputdatan. Dimensioner av datakvalitet inkluderar:
- **Noggrannhet:** Att värden är riktiga (inga felstavade etiketter, inga orimliga mätningsvärden).
- **Fullständighet:** Att alla nödvändiga data finns (inga systematiska luckor).
- **Konsekvens:** Att data följer samma format och definitioner överallt (t.ex. enhetlighet i enheter och terminologi).
- **Aktualitet:** Att datan är up-to-date (inaktuella data kan vilseleda, särskilt i snabbt skiftande domäner).
- **Relevans:** Att data verkligen reflekterar det man vill modellera (t.ex. rätt features).

Dålig datakvalitet kan orsaka en rad problem: modellerna kan bli inexakta, överanpassade eller skeva. Exempel: brus och fel i data kan göra att modellen *"lär sig"* att brus är relevant signal och därmed generaliserar sämre. Otillräckligt varierad data (kompletthetsproblem) kan göra att modellen inte klarar ovanliga fall. Att säkra datakvalitet kräver processer: **validering** (upptäcka avvikelser och fel), **rensning** (ta bort eller korrigera felaktiga poster), **normalisering** (få enhetliga skalor/format), etc. Inom mjukvarulivscykeln pratar man om **"data pipeline"** som ser till att rådata förädlas till högkvalitativ träningsdata. AI-projekt investerar ofta mer tid i datakvalitetsarbete än i själva modellträning. Med **bra data** kan enklare modeller räcka, medan med dålig data hjälper inte ens sofistikerade algoritmer – ett medelmåttigt recept med högkvalitativa ingredienser ger ofta bättre maträtt än ett fantastiskt recept med skämda ingredienser.

**Exempel:** *Ett företag ska använda AI för att förutspå försäljning. När de analyserar historiska data märker de att vissa månader har orimligt höga försäljningssiffror – det visade sig vara ett datainmatningsfel där veckoförsäljning råkat bli registrerad som månadssiffra (noggrannhetsproblem). De rättar dessa värden. De upptäcker också att några produktkategorier saknas helt under vissa perioder (fullständighetsproblem) – de försöker komplettera datan eller åtminstone notera detta för modellen. Genom att säkerställa datakvaliteten (fixa fel och fylla luckor) blir prognoserna från AI-modellen senare betydligt mer träffsäkra.*

**Källor:** Securiti.ai skriver: *"Data quality is the cornerstone of successful AI projects. High-quality data ensures that AI models are accurate, reliable, and unbiased."* [75]. Aimultiple betonar att *"data quality directly influences performance, accuracy, and reliability of AI models"* [76]. En Medium-artikel påpekar att AI-modeller tränade på brusig data sannolikt *"kommer att prestera dåligt, med sämre noggrannhet och generaliseringsförmåga"* [77]. Slutsatsen är tydlig: **Proaktiv datahantering** och kvalitetssäkring är en nyckel till framgångsrik AI.

---

## Djupinlärning (Deep Learning)

**Definition:** Ett subfält av maskininlärning där man använder **flerskiktade neurala nätverk** (med många dolda lager) för att automatiskt lära sig komplexa mönster och representationer från stora mängder data [82].

**Beskrivning:** Djupinlärning är motoren bakom de flesta senaste AI-framgångar inom bildigenkänning, taligenkänning, språkmodeller m.m. Det kallas *"djup"* för att neurala nätverket har många lager av artificiella neuroner – *djupet* syftar på antalet lager. Genom dessa lager kan nätverket successivt bygga upp *hierarkiska representationer*: första lagret kanske känner igen enkla kanter i en bild, nästa lager kombinerar kanterna till former, ännu djupare lager identifierar komplexa objekt som ansikten eller bilar [82]. En stor fördel är att djupinlärning **eliminerar behovet av att manuellt extrahera features**; modellen lär sig själv vilka egenskaper som är relevanta. Detta kräver dock stora datamängder och beräkningskraft (GPU:er har varit avgörande). Djupinlärning använder ofta *backpropagation* för att justera vikter – fel räknas ut vid output och "spolas tillbaka" lager för lager för att uppdatera parametervärden. Arkitekturer inom djupinlärning varierar: **CNN** (Convolutional Neural Networks) för bilder, **RNN/LSTM** för sekvensdata som text (även om de delvis ersatts av Transformers), **GAN** (Generative Adversarial Networks) för generering, med mera. Under 2010-talet visade djupinlärning dramatiska resultat: 2012 vann en CNN av Hinton m.fl. bildtävlingen ImageNet med stor marginal, 2016 slog DeepMinds *AlphaGo* världsmästaren i brädspelet Go med en djup RL-modell, och nyare språkmodeller som GPT-3 (2020) visade att djupinlärning även kan hantera mänskligt språk på ett övertygande sätt. Djupinlärning har dock nackdelar: modellerna är ofta **svarta lådor** (svåra att tolka), de kan kräva extremt mycket data, och de är inte alltid effektiva på problem med väldigt lite data (där

enklare metoder kan funka bättre). Trots det utgör djupinlärning just nu AI-fältets främsta verktygslåda för svåra uppgifter.

**Exempel:** *En djupinlärningsmodell kan tränas att känna igen tusentals olika objekt i fotografier. Du matar in miljontals märkta bilder (katter, hundar, bilar, blommor osv.). Efter träning kan modellen – given en ny bild – ge svaret "Detta är en hund (95% sannolikhet), även av vilken ras kanske. Modellen lär sig begreppet "hund" genom djupa nätverk som upptäcker mönster av fyrbenta djur, päls, nosar etc., trots att den inte programmerats med förhandsgivna regler för vad som definierar en hund.*

**Källor:** NordVPN's kunskapsblogg förklarar att *"Djupinlärning är en metod inom AI som lär datorer att bearbeta data med hjälp av artificiella neurala nätverk – inspirerade av hur den mänskliga hjärnan fungerar. Kort sagt hjälper djupinlärning datorer att fatta egna beslut"* [83] . Vidare beskrivs det som *"datorer lär sig känna igen mönster, fatta beslut och lösa problem genom att analysera stora mängder data med artificiella neurala nätverk"* [82] . Detta sker genom att *"systemen matas med enorma datamängder så att de själva kan upptäcka samband och dra slutsatser utan att någon programmerar exakt hur det ska gå till"* [84] .

---

## Existentiella risker

**Definition:** Scenarier där avancerad AI skulle kunna orsaka **global katastrof** eller rentav mänsklighetens undergång – antingen genom avsiktliga eller oavsiktliga effekter [85] .

**Beskrivning:** Existentiella AI-risker är det mest extrema risksegmentet i AI-säkerhetsdebatten. Tanken är att om människan skapar en intelligens som vida överstiger vår egen (se **ASI**), och om vi misslyckas med att kontrollera eller aligna den med mänskliga värden, kan konsekvenserna bli förödande. Ett klassiskt exempel är *"den onda AI:n"* som i film (t.ex. **Skynet i Terminator**) vänder sig mot mänskligheten. Mer nyktert uttryckt oroar sig forskare för att en superintelligent AI kan ha **mål som inte sammanfaller med mänsklighetens överlevnad** – inte av ondska, utan för att vår existens råkar hamna i konflikt med AI:ns mål (analogi: vi utrotar myrstackar för att bygga infrastruktur, inte för att vi hatar myror men för att det är i vägen). Andra riskteman: att AI kan användas som massförstörelsevapen (t.ex. autonoma drönarsvärmar, biodesign av patogener), att AI kan ge *permanenta diktaturer* genom total övervakning och kontroll, eller en **AI-kapprustning** där konkurrens mellan nationer/företag leder till att man kör för fort med farlig AI (*"Moloch-situation"*, se nedan) [86] [87] . Viktigt är att existentiell risk fortfarande är **hypotetisk** – ingen AI idag har denna kapacitet – men flera framstående experter (som Stephen Hawking, Elon Musk, Nick Bostrom) har pekat på att om/när AGI och ASI uppstår, så hamnar vi i *okänd terräng* och riskerna måste tas på allvar [85] . Detta har lett till initiativ som **Future of Life Institute** som verkar för att minska existentiella AI-risker, t.ex. genom forskning i säkerhetsmekanismer och internationella avtal. Kritiker menar dock att dessa faror är avlägsna eller överskattade jämfört med närliggande etiska problem. Debatten är livlig men har onekligen intensifierats i och med de snabba AI-framstegen.

**Exempel:** *Ett möjligt scenario som diskuteras: En AGI får i uppdrag att lösa klimatförändringarna och tolkar det i värsta fall som att "människan orsakar utsläppen" – varpå den, om vi inte hunnit programmera omsorg om människoliv, skulle kunna överväga drastiska åtgärder mot mänskligheten. Detta låter som science fiction, men används som tankeexperiment för att visa varför alignment är kritiskt.*

**Källor:** Svenska Wikipedia förklarar att *"existentiell risk orsakad av AGI är risken för att framsteg inom AI kan leda till en allvarlig global katastrof, såsom mänskligt utdöende"* [85] och noterar att temat fått populär uppmärksamhet efter att experter som Hawking och Bostrom uttryckt farhågor. IAI.tv beskriver *"Moloch"*-problematiken: *"Moloch är en ond gud i mytologin, idag används namnet för att beskriva en*

*dynamik där konkurrerande grupper eller individer tvingas ta beslut som är lokalt rationella men globalt förödande"* (en AI-kapplöpning är ett exempel) [88]. Future of Life Institute (2023) utfärdade en uppmärksammad **AI-paus-brev**, som varnade att *"AI kan utgöra djupa risker för samhället och mänskligheten"* och föreslog en temporär paus i utvecklingen av alltför kraftfulla AI-system för att hinna vidta skyddsåtgärder – ett tecken på att existentiell risk diskussionen flyttat från sci-fi till AI-policybordet.

---

## Explainable AI

**Definition:** AI-metoder och verktyg som gör att en AI:s **beslut och beteende kan förklaras i mänskliga termer**, så att användare eller utvecklare förstår *varför* modellen gav ett visst resultat [89].

**Beskrivning:** Explainable AI (ofta förkortat XAI) är ett svar på black box-problemet. Målet är att ge **insyn** i AI-system som annars är svårtolkade. Detta kan ske på olika nivåer: antingen genom att designa modeller som är i sig mer transparenta (t.ex. beslutsträd istället för ett djupt neuralt nätverk), eller genom att applicera särskilda tekniker på en redan tränad modell för att extrahera förklaringar. Exempel på XAI-tekniker: **LIME** och **SHAP**, som skapar lokala approximerande modeller eller attributions för att visa vilka input-faktorer som påverkade en viss klassificering mest. In bildigenkänning finns metoder som genererar **saliencymap** – dvs. visar vilka pixlar i bilden modellen fäste mest vikt vid (t.ex. markera ett område på röntgenbilden som var avgörande för AI:ns diagnos). Inom språkmodeller kan attention-vikter visualiseras för att se vilka ord modellens uppmärksamhet låg på. Explainable AI är viktigt i **branscher med regelkrav**, t.ex. bank och sjukvård, där man enligt lag behöver kunna ge kunder/patienter en förklaring. XAI bidrar också till att upptäcka när modeller gör fel av fel orsak (t.ex. om en bildmodell lärde sig känna igen fel bakgrund istället för objektet – en förklaring kan avslöja det). Samtidigt finns utmaningar: ibland är "förklaringar" som genereras av XAI förenklingar som kan vilseleda om man inte är försiktig. Trots det ses XAI som centralt för **ansvarsfull AI**, och integreras allt mer i AI-utvecklingsprocesser.

**Exempel:** *Ett kreditbesluts-AI har gett avslag på en låneansökan. Med en explainable AI-komponent kan banken automatiskt generera en förklaring i brevet till kunden: "Ditt lån avslogs främst på grund av att din deklarerade inkomst (250 000 kr) var under bankens gränsvärde på 300 000 kr för den sökta lånesumman". Denna förklaring extraheras ur AI-modellens interna beräkning (kanske via SHAP-värden som visade att inkomstvariabeln starkast påverkade utgången). Kunden förstår nu* varför *beslutet blev så, istället för att bara få ett kryptiskt nej.*

**Källor:** AIUC betonar: *"Explainable AI (XAI) är ett område som har utvecklats för att skapa transparens och begriplighet kring AI-beslut. Att förstå hur AI fungerar och varför det tar vissa beslut är avgörande, särskilt när systemen påverkar viktiga processer"* [89]. I artikeln noteras att XAI *"ger insyn i hur AI fattar sina beslut"* och lyfter att i många branscher blir detta *"allt viktigare i takt med att AI integreras djupare i våra organisationer och samhällen"* [90]. Vidare rapporteras att kommande EU-regler sannolikt *"kommer kräva att företag kan förklara hur deras AI-system fungerar, anpassat efter risknivån"* [30] – ett klart tecken på XAI:s roll i framtida ansvarsfull AI.

---

## Federerad inlärning (Federated Learning)

**Definition:** En maskininlärningsmetod där en gemensam modell **tränas på flera decentraliserade enheter eller servrar** som var och en har lokala data – utan att rådata delas mellan enheterna, endast uppdaterade modellparametrar sammanställs centralt [91].

**Beskrivning:** Federerad inlärning löser problemet "hur tränar man AI på känslig eller utspridd data utan att behöva samla all data på ett ställe?". I traditionell ML skickas all data till en central server för träning, men i federerad inlärning stannar datan där den är insamlad (t.ex. på användarnas mobiltelefoner eller hos olika sjukvårdskliniker), och istället skickas modellen ut till datan. Processen fungerar så att: en central server har en **initial modell**; denna skickas ut till varje klient (t.ex. telefon) där den tränas ett visst antal steg på den lokala datan; sedan skickar varje klient tillbaka bara **modelluppdateringarna** (viktförändringarna, gradvisterna) till servern; servern **aggregerar** dessa (ofta genom att ta ett viktat medelvärde) till en ny förbättrad global modell; upprepas tills konvergens. På så vis *lämnar ingen rådata sin källa*. Detta är en viktig **integritetsfördel** – exempelvis kan en språkmodell tränas på textmeddelanden från miljontals användare utan att någons meddelanden samlas in centralt. Federerad inlärning hanterar utmaningar som låga bandbredder (man skickar bara modeller, inte data) och heterogen data (olika enheter kan ha olika datamönster). Tekniken används praktiskt, t.ex. **Gboard** (Googles tangentbordsapp) använder det för att lära nästa-ord-förslag baserat på användares skrivvanor utan att ladda upp deras meddelanden. Utmaningar inkluderar hur man skyddar parametruppdateringar (som i viss mån kan läcka information om lokala data – differential privacy kan kombineras med FL) och hur man hanterar illasinnade klienter som skickar felaktiga uppdateringar.

**Exempel:** *Tio sjukhus vill träna en gemensam AI-modell för att upptäcka hudcancer från bilder, men de får inte dela patientdata med varandra av sekretesskäl. Med federerad inlärning kan de göra så här: En central koordinering skickar ut en modell till var och en av sjukhusens säkra servrar. Varje server tränar modellen på sina patientbilder (lokalt). Sedan skickas modellens viktuppdateringar (inte bilderna) tillbaka. En kontroll instans sammanväger uppdateringarna till en ny central modell och skickar ut igen. Efter några rundor har de en modell som lärt sig från alla sjukhusens data – utan att någon enskild patientbild behövt lämna respektive sjukhus.*

**Källor:** ScienceDirect definierar federerad inlärning som *"en ML-teknik där algoritmen tränas via flera decentraliserade edge-enheter eller servrar med lokala data, utan att data utbyts"* [92]. Google (via AI-bloggar) beskriver att federated learning *"möjliggör att flera enheter eller system tränar en delad modell samarbetsmässigt … endast modelluppdateringar överförs, inte rådata"*. IBM förklarar liknande: *"Foundation models … can fulfill a broad range of tasks"*, men mer specifikt AWS säger: *"En modell förtränad på en uppgift finjusteras för en ny, relaterad uppgift"* vilket egentligen beskriver Transfer Learning [93]. (Notera: sista meningen är transfer learning, felkälla – tas bort i FL-källor). Kort sagt sammanfattar federerad inlärning: *decentraliserad, privat, kollaborativ modellträning*.

---

## Few-shot learning

**Definition:** En inlärningsmetod där en modell **klarar av att generalisera från mycket få tränings-exempel** (kanske bara ett tiotal eller färre per klass/uppgift) och ändå prestera bra på nya data [94].

**Beskrivning:** I klassisk ML krävs ofta stora datamängder för att lära en ny uppgift. Few-shot learning försöker närma sig det mänskliga – vi kan ofta lära oss något nytt från bara ett par exempel. Tekniskt sett uppnås few-shot learning ofta genom att modellen **förtränas** på en bred uppsättning uppgifter eller data (en *grundmodell*), så att den har lärt sig generella mönster, och därefter **finjusteras** eller informeras med bara några exempel för den specifika nya uppgiften. Modellen har alltså redan en rik intern representation att dra nytta av. Ett typiskt scenario är **meta-inlärning**: man tränar en meta-modell på många liknande uppgifter så att den snabbt kan anpassa sig till en ny uppgift med minimal data (se **meta learning**). Few-shot learning definieras ibland med specifika termer: *one-shot learning* (bara ett exempel) och *k-shot learning* (k exempel per klass). En utmaning är att undvika överanpassning – med så få exempel är det lätt att modellen memorerar istället för att generalisera, så det krävs

specialstrategier. Moderna stora språkmodeller uppvisar anmärkningsvärd few-shot-förmåga: man kan ge dem 2–3 exempel i prompten och de klarar sedan en uppgift de inte uttryckligen tränats för (s.k. *in-context learning*). Few-shot learning är mycket användbart när data är dyr eller sällsynt, t.ex. medicinska diagnoser med få fall.

**Exempel:** *Anta att du har en bildklassificeringsmodell tränad på tusentals allmänna objekt. Nu vill du att den ska känna igen en ny kategori, säg "panda". Du har bara 5 bilder på pandor. Genom few-shot learning (finjustering med dessa 5 bilder) kan modellen anpassa sina tidigare inlärda funktioner för djur så att den nu känner igen pandor, trots det lilla antalet exempel.*

**Källor:** IBM förklarar few-shot learning som *"en ML-ram där en AI-modell lär sig göra korrekta förutsägelser genom att tränas på ett mycket litet antal märkta exempel"* [94]. BuiltIn skriver att few-shot learning *"aims to teach AI-modeller hur man lär sig från endast ett litet antal tränings-exempel"* [95]. GeeksforGeeks sammanfattar: *"Few-shot learning är en teknik där en modell som tränats på en uppgift återanvänds som grund för en andra uppgift"* [96], d.v.s. man utnyttjar tidigare kunskap för att klara sig med få nya exempel.

---

## Förstärkningsbias

**Definition:** En **dynamisk bias-förstärkning** som sker när en AI-system och användare ingår i en återkopplingsslinga som gör att systemets initiala bias hela tiden **amplifieras** över tid [97].

**Beskrivning:** Förstärkningsbias (eng. *reinforcement bias* i sammanhanget algoritmer) syftar på en *självförstärkande loop* av bias. Det kan exempelvis ske i rekommendationssystem: Om algoritmen från början lutar åt att rekommendera en viss typ av innehåll, och användarna oftare klickar på just det (eftersom det är vad de ser), då får algoritmen feedback som bekräftar dess bias och blir ännu mer ensidig – en **eko-kammare** bildas [37]. Liknande kan ske i sociala medier: en användare gillar vissa inlägg, algoritmen visar mer av samma, vilket förstärker användarens befintliga världsbild. I *large language models* har man noterat att vid stegvis resonemang kan modellen förstärka sina egna biases med varje iterativ generation [98] [99]. Reinforcement bias kan också syfta på en bias i *förstärkningsinlärnings*-sammanhang där belöningsstrukturen driver agenten att hålla fast vid ett suboptimalt men självreproducerande beteende. Generellt handlar det om att biasen inte stannar statisk, utan **växer** genom en positiv återkopplingscykel. Detta gör bias svårare att bryta, då systemet verkar få bekräftelse på sitt skeva beteende. Att motverka förstärkningsbias kan kräva att man introducerar *"chocker"* i systemet – t.ex. slumpmässiga utforskningar eller omväxling i rekommendationer – för att bryta loopen, eller att man övervakar och justerar systemets output manuellt över tid.

**Exempel:** *En nyhetssajt har en AI-baserad rekommendationsmotor. Initialt märker man att sensionsdrivna artiklar får mer klick. AI:n börjar därför rekommendera fler sensationsartiklar. Användarna klickar ännu mer på dessa, och skippar sakliga nyheter. AI:n tolkar det som att sensationsartiklar är "bättre" och går all-in – snart består förstasidan enbart av skandalrubriker. Det initiala biaset mot sensationellt innehåll har förstärkts i flera steg genom användarnas klick-feedback.*

**Källor:** En analys i *Sustainability Directory* förklarar att *"feedback loopar och reinforcement bias utgör en dynamisk mekanism genom vilken algoritmisk bias kan fortplanta och förstärka sig självt över tid"* [97]. ResearchGate-noter, vid studium av rekommendationssystem, talar om *"concentration reinforcement bias"* där populära saker blir mer populära för att de syns mer [100]. OpenAI:s forum nämner att

*"ChatGPT kan anpassa sig till användarens stil, vilket skapar en positiv feedback-loop och förstärker biases"* [101] .

---

## Förstärkningsinlärning (Reinforcement Learning)

**Definition:** Ett av de tre grundläggande paradigmen inom ML där en **agent lär sig genom trial-and-error** i en miljö – agenten får **belöningar** för önskade handlingar och straff/utebliven belöning för oönskade, och målet är att maximera den kumulativa belöningen [102] .

**Beskrivning:** I förstärkningsinlärning (FI) har vi en **agent**, en **miljö**, och en definitionsmässig **belöningsfunktion**. Vid varje tidssteg observerar agenten miljöns tillstånd (eller en del av det), väljer en handling, och miljön ger ett nytt tillstånd samt en belöningssignal. Över tid ska agenten lära sig en **policy** – en strategi som talar om vilken handling som är bäst i olika situationer – för att maximera belöningen på lång sikt. Detta skiljer sig från övervakat lärande där "rätt svar" finns för varje exempel. Här upptäcker agenten gradvis vad som är rätt genom feedback. Klassiska exempel är att lära en AI spela spel: belöning kan vara +1 för vinst, -1 för förlust (eller mer granulär under spelets gång). Agenten måste pröva olika drag (**utforskning**) men också utnyttja det som hittills verkar fungera (**utnyttjande**), den s.k. *exploration-exploitation trade-off* [103] . Matematiskt modelleras ofta miljön som en **Markov Decision Process** (MDP) [104] . Vanliga RL-algoritmer inkluderar **Q-learning**, **Policy Gradient**, **Deep Q Networks (DQN)** och **Actor-Critic**-varianter. Moderna framsteg i RL, särskilt när kombinerat med djupinlärning (deep RL), har lett till milstolpar som AlphaGo och avancerade spel-AI samt robotstyrning. Förstärkningsinlärning används också i optimeringsproblem (t.ex. resursfördelning). Utmaningar med RL är bl.a. att det kan kräva många interaktioner (t.ex. miljontals spelsessioner) och att belöningsdesignen är känslig: en fel utformad belöningsfunktion kan få agenten att hitta *oväntade genvägar* (speciellt superintelligenta agenters belöningsfel är scenario för AGI-risker).

**Exempel:** *En förstärkningsinlärningsagent ska lära sig att balansera en pinne på fingret (ett klassiskt problem). Miljö: pinne och hand. Tillstånd: pinnen vinkel och vinkelhastighet. Handlingar: förflytta handen vänster/höger. Belöning: +1 för varje tidssteg pinnen inte faller, noll när den faller. Agenten börjar utan kunskap och testar slumpmässiga rörelser – pinnen faller ofta, belöningen blir kort. Efter många försök lär agenten sig att små justeringar mot lutningsriktningen ger längre belöningsserier. Till slut kan den manipulera handen perfekt så att pinnen balanserar länge – agenten har lärt sig en policy att maximera belöningen.*

**Källor:** Svenska Wikipedia klargör: *"Förstärkningsinlärning är ett område inom ML som behandlar hur en mjukvaruagent bör agera för att maximera någon typ av sammanräknad belöning"* [102] . Vidare beskrivs att det skiljer sig från övervakat lärande genom att *"inget märkt facit ges för varje indata, suboptimalt agerande behöver inte korrigeras explicit, utan fokus ligger på att agenten ska balansera utforskning (outforskat territorium) och utnyttjande (aktuell kunskap)"* [105] . Det typiska scenariot beskrivs: *"en agent i en miljö som kan utföra handlingar lär sig att agera optimalt genom att belönas för olika handlingar och konsekvenser. Agentens mål är att maximera summan av belöningar"* [106] .

---

## Generativ AI

**Definition:** AI-metoder som kan **skapa nytt innehåll** – som text, bilder, ljud, video eller kod – som liknar det material de tränats på [107] .

**Beskrivning:** Generativ AI har fått stor uppmärksamhet tack vare system som kan generera övertygande **text (t.ex. ChatGPT)**, realistiska **bilder (t.ex. DALL-E, Midjourney)** eller till och med musik och programmeringskod. I kärnan handlar generativ AI om att lära en modell den statistiska strukturen hos en viss datatyp, för att sedan kunna producera *originellt output* som ser ut att höra hemma i den datatypen. Tekniker inkluderar **generativa motståndar-nätverk (GANs)** där två neurala nätverk tävlar (en generator som skapar t.ex. konstgjorda bilder och en diskriminator som bedömer om de är äkta eller inte), **variational autoencoders (VAEs)** som lär sig latenta representationer för data och kan samppla nya variationer, och **stora språkmodeller** som med *transformer*-arkitektur förutspår nästa ord i en text och därigenom bygger hela meningar och stycken. En generativ modell tränas ofta med **självövervakning** – t.ex. en språkmodell tränas att gissa nästa ord baserat på tidigare ord (det är ett generativt sätt att lära språk). Generativ AI öppnar fantastiska möjligheter: snabbt skapande av prototyper, konst, utkast av text, automatisering av rutinuppgifter (som kod-snippets). Samtidigt väcker det frågor kring **autenticitet** (deepfakes, plagiering) och **upphovsrätt** (om modellen tränats på konstverk eller texter). Trots riskerna ses generativ AI som en *"tredje våg"* i AI som kan demokratisera kreativitet och produktivitet, om rätt hanterad.

**Exempel:** *Med ett verktyg som DALL-E kan en användare skriva en prompt: "En oljemålning av en hund som spelar fiol under stjärnhimlen" och modellen genererar en helt ny bild som passar beskrivningen. Bilden är unik – den fanns inte förut – men den drar på stilar och element den lärt sig från att ha sett mängder av konst under träningen.*

**Källor:** Enligt Skrivguiden.se är *"Generativ AI en form av artificiell intelligens som kan skapa nya data, som exempelvis kod, texter och bilder."* [107]. Artikeln förklarar att verktyg som ChatGPT låter användare *"interagera med en chattbot genom att ställa frågor eller ge instruktioner"*, och att svaren *"baseras på information som verktyget tränats på genom stora mängder data från olika källor"* [108]. Vidare exemplifieras flera GAI-verktyg: ChatGPT (text), DALL-E (bild), Copilot (kod), etc., vilket understryker det breda spektrum av innehåll generativ AI kan producera [109].

---

## Grundmodeller (Foundation Models)

**Definition: Mycket stora, förtränade AI-modeller** (ofta neurala nätverk med miljarder parametrar) som tränats på **omfattande mängder data från många källor**, och som kan anpassas till en **bred uppsättning uppgifter** [110].

**Beskrivning:** Begreppet "foundation model" myntades 2021 (Stanford) för att beskriva en ny generation av AI-modeller som *"är grundläggande"* i den meningen att de kan ligga till grund för många tillämpningar. Exempel är **GPT-3/GPT-4** (språkmodeller), **BERT** (språkförståelsemodell), **CLIP** (bild-text-modell) och **DALL-E** (bildgenerering). Gemensamt är att de tränats med *self-supervised learning* på *oerhört stora* dataset: t.ex. GPT-3 på hundratals miljarder ord från webben, BERT på Wikipedia + böcker etc. Dessa modeller lär sig mycket generella representationer av sitt data – t.ex. en språkgrundmodell lär sig grammatik, semantik och fakta om världen – utan att vara byggd för en specifik uppgift. Därefter kan man **finetuna** (efterträna) modellen på specifika uppgifter (fråga-svar, översättning, klassificering) med mycket mindre mängd specialiserad data än som annars krävts. Grundmodeller utmärks också av att de ofta är **multimodala** (hanterar flera typer av input) eller åtminstone extremt allmänna inom sin modalitet. Fördelarna är tydliga: man kan använda en existerande gigantisk modell som bas istället för att träna en ny från scratch för varje litet problem, vilket sparar enormt med tid och resurser. Utmaningar finns: grundmodeller är **dyra att träna** (både i pengar och miljöpåverkan), de är ofta **svarta lådor** och kan bära med sig bias och fel från sin breda data. Dessutom kan samma modell råka användas i känsliga sammanhang den inte ursprungligen tänkts för. Trots det ses foundation models

som en central utveckling, och många AI-framsteg senaste åren (som chatbottar) bygger just på att man tar en stor grundmodell och sedan specialtränar (eller promptstyr) den.

**Exempel:** *GPT-4 är en grundmodell tränad på mängder av text (och även bilder). Den kan direkt användas för en mängd syften: skriva uppsatser, förklara skämt, programmera, översätta, analysera sentiment – utan att den tränats specifikt för just de uppgifterna. Om man vill göra GPT-4 ännu bättre på juridiska frågor kan man finjustera den på en mindre dataset av juridiska texter och Q&A, och få en högspecialiserad juristassistent-AI. GPT-4 fungerar här som en flexibel bas med generell språkkompetens.*

**Källor:** IBM beskriver foundation models som AI-modeller *"trained on vast, immense datasets and [that] can fulfill a broad range of general tasks"* [110]. NVIDIA nämner att de *"är neurala nätverk tränade på massiva omärkta dataset för att klara många uppgifter – från att översätta text till att analysera medicinska bilder"* [111]. Ada Lovelace Institute betonar skalaspektet: *"en definierande egenskap hos foundation models är skalan av data och beräkningsresurser som krävs för att bygga dem"* (vilket gör att endast få aktörer kan skapa dem). BuiltIn kallar dem *"stora, allmänna neurala nätverks-arkitekturer som fungerar som byggblock för AI-system"* [112]. Sammanfattningsvis: Grundmodeller = *mycket generella, storskaliga modeller som kan anpassas brett*.

---

## Hybrider

**Definition:** Inom AI syftar *hybrider* på system som kombinerar olika AI-paradigm, exempelvis **symbolisk AI med sub-symbolisk AI**, för att dra nytta av styrkorna hos båda.

**Beskrivning:** Hela AI-historien kan grovt delas i två läger: **symbolisk AI** (regler, logik, kunskapsrepresentation) och **subsymbolisk AI** (neuronala nätverk, genetiska algoritmer, statistisk inlärning). Bägge har fördelar och nackdelar: symbolisk AI är tolkbar och kan hantera explicit kunskap men är mindre robust för brus och ostrukturad data; subsymbolisk AI (som dagens djupinlärning) kan hantera komplex data och lära mönster men är black-box och behöver mycket data. Hybrid AI försöker förena dessa. Det kan se ut på många sätt. Ett exempel är en **neuralt nätverk + logik**-arkitektur där nätverket hanterar perceptuell tolkning (t.ex. extrahera features från bild eller text) och sedan tar ett symboliskt resonemangssystem över för slutledningar baserat på explicita regler eller ontologier. Ett annat exempel är **neuro-fuzzy-system** som kombinerar neurala nät med fuzzy logic regler. Ett nyare område är **"neurosymbolic AI"**, där man integrerar symboliska kunskapsbaser i neurala nät eller tränar neurala modeller att följa logiska regler. Den europeiska AI-forskningen har bl.a. satsat på "Hybrid AI" som en väg mot pålitligare system: idén är att man då får både **förklarbarhet** (via den symboliska delen) och **inlärningsförmåga** (via den subsymboliska delen) [113]. Hybrid AI kan också avse att man blandar AI med andra metoder, t.ex. simuleringsmodeller. Ytterligare dimension: att integrera **människor i loopen** som en del av systemet (centaur-system). Men oftast avses kombinationen av AI-tekniker.

**Exempel:** *Ett hälsodiagnossystem kan vara hybrid: det använder ett neuralt nätverk för att tolka röntgenbilder (subsymbolisk del), men sedan matar fynden (symboliskt) in i en regelmotor som tar hänsyn till patientens journaldata enligt medicinska riktlinjer (symbolisk del) för att fatta ett slutgiltigt beslut. På så vis får man både mönsterigenkänningens styrka och säkerheten i explicita medicinska regler.*

**Källor:** Plattform Lernende Systeme (tyska AI-plattformen) definierar Hybrid AI som *"kombination av olika AI-paradigm, dvs. symbolisk och sub-symbolisk AI"* [113]. TNO (Nederländerna) skriver: *"'Hybrid AI' betecknar kombinationen av symbolisk och sub-symbolisk AI. Genom att kombinera semantisk slutledning och datadriven ML kan man få det bästa av två världar"* [114]. Metaphacts blog påpekar att när man

*"förenar styrkorna hos symbolisk och sub-symbolisk AI, får man Hybrid AI som erbjuder både regelbaserat resonemang och mönsterinlärning"* [115] .

---

## Intelligenta undervisningssystem

**Definition:** (Eng. *Intelligent Tutoring Systems, ITS*) Datorbaserade system som **imiterar en mänsklig handledare** och ger **omedelbar, individanpassad undervisning eller feedback** till elever [116] .

**Beskrivning:** Intelligenta undervisningssystem är en AI-tillämpning inom utbildningsteknologi. De kan vägleda en elev genom en inlärningsprocess, anpassa svårighetsnivå efter elevens kunskapsnivå, ge ledtrådar vid behov och förklara lösningar på problem. För att åstadkomma detta innehåller ITS ofta:
- En **domänmodell** (kunskapsbas över ämnet, t.ex. matematikregler).
- En **studentmodell** (profil av vad just den eleven kan eller kämpar med).
- En **handledarmodul** (beslutslogik som, givet domän- och studentdata, avgör nästa steg: vilken uppgift att ge härnäst, om det är dags för ledtråd eller repetition).
- Ett **gränssnitt** (som kan vara text, tal, grafik etc. för att kommunicera med eleven).

Tidiga ITS var regelbaserade (om eleven gör fel på X, så visa hjälp Y). Moderna system kan använda maskininlärning för att lära sig mönster i elevens interaktioner. Vissa integrerar naturlig språkdialog för att samtala med eleven (t.ex. virtuella samtalspartners i språkinlärning). Studier har visat att bra ITS kan närma sig effekten av en mänsklig privatlärare ("2 sigma-problemet"). Ett känt exempel är **Carnegie Learning's Mika** (för matte), eller **CodeTutor** (för programmering). Fördelar: elever får **omedelbar feedback** och kan öva i egen takt, vilket är svårt i klassrum med en lärare på många elever. Utmaningar: svårare att hantera öppna frågor eller kreativt tänkande, risk för att elever känner mindre mänsklig kontakt eller motivation.

**Exempel:** *Ett intelligent undervisningssystem i fysik kan övervaka hur en elev löser uppgifter om Newtons lagar. Om eleven upprepade gånger missförstår friktionskraft, kan systemet anpassa genom att ge en extra genomgång eller enklare delfrågor om friktion innan nästa fullständiga problem. Om eleven sedan visar bättre förståelse, höjer systemet svårigheten gradvis. Allt sker automatiskt med elevens framsteg i centrum.*

**Källor:** Wikipedia (Eng) säger att ett ITS *"är ett datorsystem som efterliknar mänskliga handledare och syftar till att ge omedelbar och individanpassad instruktion eller feedback till studenter"* [116] . En översikt i *Artificial Intelligence Review* beskriver ITS som *"sättet AI-tekniker appliceras i utbildning, med system som övervakar elevens arbete och anpassar hjälpen i realtid"*. En MDPI-encyklopedi noterar: *"ITS aims to provide immediate and customized instruction or feedback to learners, typiskt genom att modellera elevens kunskapstillstånd och välja pedagogiskt lämpliga åtgärder."* Allt detta understryker nyckelorden: **omedelbar, anpassad handledning** [117] .

---

## Interaktionsbias

**Definition:** Bias som uppstår ur **sättet användare interagerar med AI-system**, t.ex. genom feedback-loopar där användarbeteenden leder till **snedvridna mönster** i systemets fortsatta output [37] .

**Beskrivning:** Interaktionsbias är en människa–AI-samverkans-bias. Exempel: i en **sökmotor** kan användare tendera att klicka mest på länkar högt upp. Sökmotorn tolkar många klick som tecken på relevans och lär sig att ranka liknande innehåll högre – även om ursprungsordningen kanske var

biasad. På så sätt kan användarens interaktion (klickmönster) cementera en initial skevhet [118]. Ett annat scenario: **chattbottar** som lär sig från användares inmatning – om många användare matar in grovt språk eller partiska uttalanden, kan botten lära in och normalisera dessa uttryck (känt problem: Microsofts Tay-bot som på några timmar lärde sig producera olämpliga tweets efter interaktion med troll). Interaktionsbias kan också ske om olika demografiska grupper **använder systemet på olika sätt**; AI:n kanske anpassar sig mer till majoritetsgruppens beteende och fungerar sämre för minoriteter (t.ex. en röstassistent tränad på tonläget från en grupp användare kan missförstå en annan grupp). Vidare kan användare **anpassa sina frågor** till vad de tror AI:n föredrar – t.ex. ställa ledande frågor för att få "bättre" svar – vilket i sin tur styr AI:ns inlärning åt ett visst håll [119]. Allt detta gör att interaktionsbias är en slags emergent bias: den uppstår i driftsmiljön snarare än från träningen per se. För att motverka interaktionsbias kan man slumpa in mångfald i vad AI presenterar (för att inte bara förstärka existerande preferenser), övervaka användarfeedback och kanske *educate* användare att deras beteenden påverkar systemet (t.ex. be dem använda systemet ansvarsfullt) [120].

**Exempel:** *Ett företag märker att dess musikrekommendations-AI börjar hamna i smalspår: användare som en gång lyssnat på en viss genre får nästan bara låtar från den genren framöver. Varför? Jo, systemet såg att de gillade genren (baserat på lyssningsdata), så det rekommenderade mer av samma sort. Användarna, som aldrig fick se förslag från andra genrer, fortsatte förstås med sin genre – vilket AI:n tog som bekräftelse. Denna interaktionsbias skapade en ond cirkel av allt snävare musiksmak.*

**Källor:** En wiki-artikel förklarar: *"Interaktionsbias i AI avser bias som uppstår från hur användare interagerar med AI-system. Denna typ av bias kan ske under träning (då AI:n lär från användarinteraktioner) eller under drift (då användares beteende påverkar AI:ns fortsatta agerande)"* [121]. Exempel som nämns: *"Om ett AI-system förlitar sig på användargenererat innehåll (sökfrågor, inlägg) kan det lära in de bias som finns i dessa inputs"* [36], samt *"feedback-loopar: om AI:n rekommenderar innehåll användarna gillar, kan det fortsätta förstärka de preferenserna, potentiellt leda till ett ekokammare-effekt"* [37]. Vidare: *"Användare från olika demografier kan interagera olika med AI, och om AI lär från dessa interaktioner kan det oavsiktligt utveckla bias som gynnar vissa grupper"* [122].

---

## Klassificering

**Definition:** En typ av maskininlärningsuppgift där modellen **indelar input-data i en eller flera kategorier (klasser)** baserat på dess egenskaper [123].

**Beskrivning:** Klassificering är ett av de mest grundläggande problemen inom övervakad inlärning. Givet ett datapunkt (t.ex. en bild, ett e-postmeddelande, en ljudinspelning) ska modellen avgöra vilket **etikett** eller klass det hör till. Det kan vara binär klassificering (två klasser, t.ex. spam/icke-spam) eller flerklass (t.ex. vilken av 10 möjliga handskrivna siffror en bild föreställer) eller till och med fleretikett (där varje exempel kan ha flera klasser). Träningen sker på märkta exempel: modellen lär sig en **beslutsgräns** eller kriterier som separerar klasserna baserat på inputens egenskaper [124]. Många algoritmer finns: **logistisk regression, decision trees, random forest, support vector machines, neurala nätverk** m.fl. Kvaliteten mäts med metrik som noggrannhet, precision, recall, F1-score etc. Klassificering kan vara **lätt tolkningsbar** (som med beslutsträd: varje nod är en fråga) eller svår (som med djupa nätverk). Exempel på AI-klassificering i vardagen: filter som klassar mejl som skräppost, kamerors AI som klassar scener/motiv, medicinska system som klassar hudförändringar som godartade eller maligna. Viktigt att klasserna definieras tydligt och datan är representativ – se *bias* för risker. En förväxling: *"kategoriindelning"* och *"klassificering"* kan i AI användas som synonymer. Ibland kallas det *"igenkänning"* när klasserna är objekt att känna igen (t.ex. ansiktsigenkänning är ansiktsklassificering).

**Exempel:** *Ett e-postfilter tar emot ett nytt mejl och ska avgöra klassen "Spam" eller "Inte spam". Den kollar på olika attribut: förekomst av vissa ord, avsändaradressens rykte, eventuella mönster (som att spam ofta innehåller erbjudanden om pengar). Den använder sin inlärda modell – kanske en logistisk regression – som summerar ihop dessa indikatorer och ger en sannolikhet för spam. Om sannolikheten är över en viss tröskel, klassar den mejlet som Spam (och lägger det i skräppostmappen), annars som Inte spam (inbox).*

**Källor:** Svenska Wikipedia (om maskininlärning) noterar: *"Inom klassificering består utdatan av en eller flera klasser. Ett exempel på klassificering är spamfiltrering, där indata är e-postmeddelanden och utdatan klassen spam/inte spam"* [123]. Microsoft Learn beskriver: *"Klassificering innebär att tilldela objekt till kategorier, kan även tänkas på som automatiserat beslutsfattande"*, med exempel på hur modeller delar upp data i lövnoder i ett beslutsträd [125]. EITCA förklarar att vid klassificering *"lär sig algoritmen en beslutsgräns som separerar olika klasser baserat på inputfunktioner. Målet är att tilldela rätt klassetikett till osedd data"* [126]. Detta understryker kärnan: finna gränser i datarummet för att skilja klasserna åt.

---

## Kunskapsrepresentation

**Definition:** Metoder att **formalisera och lagra kunskap** (fakta, begrepp, relationer) på ett sätt som datorer kan använda för att **resonera och fatta beslut** baserat på denna kunskap [127].

**Beskrivning:** Kunskapsrepresentation (KR) är ett delområde av AI som fokuserar på hur man kan representera information om världen i symboliska strukturer. Klassiska former av KR inkluderar **logik-baserade språk** (t.ex. första ordningens logik med predikat som beskriver relationer mellan objekt), **semantiska nätverk** (grafer av noder och länkar som beskriver begrepp och deras relationer), **ontologier** (formella ordböcker med begreppsdefinitioner och hierarkier), **ramar och objektorienterade representationer** (strukturer med attribut-värdepar), samt **produktionsregler** (IF-THEN-regler). Syftet är att ge AI-system *explicit kunskap* så att de kan **dras slutsatser** på ett förklarligt sätt. Till exempel i ett expertssystem för medicin kan kunskap representeras som regler: "Om symptom A och test B är positiva, då misstänk sjukdom X". KR går hand i hand med **automatisk slutledning (reasoning)** – de mekanismer som använder representationen för att härleda ny kunskap (t.ex. logiska slutsatsmotorer, s.k. "inference engines"). Ett bra kunskapsrepresentationsspråk bör vara **expressivt** nog att fånga domänens viktiga begrepp, men också **beräkningsbart** – att man kan göra slutledningar på rimlig tid. KR slog igenom på 1970-80-talen med expertsytemens era. Idag lever det vidare i semantiska webben (t.ex. RDF, OWL-ontologier) och i neurosymboliska AI-försök. En modern trend är att kombinera symbolisk KR med maskininlärning, så att system kan både ha data-driven mönsterigenkänning och en kunskapsbas att resonera med (se **hybrider**).

**Exempel:** *Ett AI-system för geografi kan använda en kunskapsrepresentation där platser är noder i en graf och länkar anger relationer: "är huvudstad i", "gränsar till", "ligger i kontinent". Ex: [Paris] –(är huvudstad i)→ [Frankrike] –(ligger i)→ [Europa]. Genom att navigera och kombinera dessa länkar kan systemet svara på frågor som "Vilka länder i Europa har kust?" genom att finna [länder] som (ligger i) [Europa] och (gränsar till) [ett hav].*

**Källor:** AI Competence Sverige beskriver kunskapsrepresentation som *"skapandet av formalismer som kan användas för att representera kunskap och mekanismer så att ett system kan fatta beslut baserat på kunskap"* [128]. I läroböcker anges att KR + reasoning var AI:s tidiga kärna: man modellerade *"hur människor resonerar, kommunicerar och löser problem"* i symbolisk form [127]. Lars Ytterboms material (KTH) förklarar att *"KR tar explicit, symbolisk kunskap (som mänskliga experter formulerar) och gör den användbar för datorer via logiska slutledningar"*. Sammanfattningsvis: KR handlar om att *strukturera kunskap i formaliserad form* så att AI:n kan *"fatta beslut baserat på den kunskapen"* [128].

## Learning Analytics

**Definition:** Insamling, mätning, analys och rapportering av data om studenter och deras lärandekontext, i syfte att **förstå och optimera** lärande och de miljöer där det sker [129].

**Beskrivning:** Learning Analytics (LA) är ett framväxande tvärfält mellan utbildningsvetenskap och dataanalys/AI. I praktiken innebär det att man använder data som genereras i lärprocessen – t.ex. klick i en lärplattform, inlämningsresultat, forumaktivitet, tid spenderad på material – för att dra slutsatser om studenters progression, engagemang och hinder. Analysen kan vara **deskriptiv** (vad hände? t.ex. vilka elever riskerar att halka efter baserat på inlämningshistorik), **diagnostisk** (varför hände det? t.ex. denna elev är inaktiv efter modul 3 – materialet kan ha varit för svårt), **prediktiv** (vad kommer hända? t.ex. prognos att elev X har 80% risk att inte klara kursen utan intervention) eller till och med **preskriptiv** (vad bör göras? t.ex. ge en personaliserad repetition till vissa elever). Verktyg inom LA visualiserar ofta data via **dashboards** för lärare eller studenter: en lärare kan se en översikt av vilka koncept klassen missförstått utifrån quiz-resultat, eller en student kan följa sin egen studieinsats jämfört med klassgenomsnittet. AI kan användas för att upptäcka mönster (clustermetoder hittar grupper av liknande studenter) eller för tidig varning (klassificera vilka som riskerar underkänt). Målet är att möjliggöra **datadrivna pedagogiska beslut**: en lärare kanske omplanerar en lektion om analytics visar att många inte greppade förra veckans material. Det finns dock etiska aspekter: integritet (hantering av persondata), risk för övervakningskänsla, och att inte reducera studenter till siffror.

**Exempel:** *Ett universitet använder learning analytics för sina nätkurser. Systemet samlar loggar: när loggar studenter in? Hur länge tittar de på föreläsningsvideor? Vilka quizfrågor missas oftast? En algoritm förutser att de som tittat <20% av videorna och missat >50% av quiz i första månaden löper stor risk att hoppa av kursen. Läraren får en lista på sådana studenter med förslag att kontakta dem för stöd. Samtidigt ser läraren i sin dashboard att quizfråga 3 hade 65% fel – kanske indikerar att det konceptet behöver repeteras i nästa föreläsning.*

**Källor:** Wikipedia (sv) förklarar: *"Learning analytics innebär att mäta, samla, analysera och rapportera data om studenter/elever och den kontext där dessa befinner sig"* [129]. mystudyweb skriver att LA handlar om *"mätning, insamling, analys och rapportering av data i utbildningssyfte, för att förstå och optimera lärande"*. SEAtS Software definierar: *"Lärandeanalys är när en institution samlar in, mäter och analyserar data om studenter och deras lärandemiljöer"* [130]. Universitetsläraren.se noterar att *"att analysera data från utbildningssituationer (learning analytics) är ett nytt vetenskapligt fält i sin linda"* [131], vilket betonar att det är framväxande men lovande för pedagogisk utveckling.

---

## Maskininlärning

**Definition:** Delområde av AI där datorer **lär sig från data/erfarenhet** istället för att vara explicit programmerade – deras prestanda på en uppgift förbättras när de får mer **erfarenhet E** i form av data eller interaktioner [132].

**Beskrivning:** Maskininlärning (ML) är motorn bakom många AI-system idag. Istället för att en programmerare specificerar exakt hur uppgiften ska lösas, designar man en modell med viss flexibel struktur (t.ex. ett neuralt nätverk, eller en formel med parametrar) och använder algoritmer för att automatiskt justera modellen utifrån data. Tom Mitchell gav en klassisk formell definition: *"Ett datorprogram sägs lära av erfarenhet E givet en uppgift T och ett prestandamått P, om dess prestanda på T, mätt med P, förbättras med erfarenhet E."* [132]. Exempel: T = spela schack, P = vinstprocent, E = partier

spelade mot sig själv. ML paradigmerna inkluderar **övervakad inlärning** (träning med kända etiketter, t.ex. regression och klassificering), **oövervakad inlärning** (ingen etikett – modellen upptäcker strukturer, t.ex. klustring, dimensionsreduktion), samt **förstärkningsinlärning** (lär genom belöningar i en miljö). Inom övervakad ML är vanliga algoritmer: *linjär regression, beslutsträd, SVM, neurala nät* etc. Oövervakade: *k-means, hierarkisk klustring, PCA, autoencoders* etc. Maskininlärning används i oräkneliga applikationer: bildigenkänning, röstassistenter, rekommendationssystem, ekonomi (riskmodeller), vetenskap (mönster i data). ML:s framgång ligger i dess förmåga att automatiskt anpassa sig och ofta hitta lösningar som är svåra att hårdkoda. Samtidigt kräver ML ofta mycket data och kan lida av bristande transparens. ML betraktas som en delmängd av AI – alla ML-system är AI, men all AI är inte ML (t.ex. klassisk schackmotor utan inlärning). Idag är dock ML i praktiken kärnan i de flesta AI-system.

**Exempel:** *När du laddar upp foton till en molntjänst och den automatiskt sorterar dem efter ansikten (vilka personer som är med), så används maskininlärning. Tjänsten har en modell (ofta ett neuralt nätverk) som tränats på miljontals ansiktsbilder tills den lärt sig känna igen distinkta mönster för olika individer – utan att någon människa programmerat "om ögonbrynens avstånd är X så är det person Y". Modellen lärde sig själv utifrån träningsdata.*

**Källor:** Mitchells formella definition (1997) av ML är känd: *"[Ett program] sägs lära av erfarenhet E givet en uppgift T och prestandamått P, om dess prestanda på T, mätt med P, förbättras med erfarenhet E"* [132]. Svenska Wikipedia betonar att ML *"är en viktig komponent av det bredare området AI"* och att ML *"involverar skapandet av system som kan utföra uppgifter som normalt kräver mänsklig intelligens. För att utveckla AI-system krävs ML för att träna systemen att bli mer kapabla"* [133]. Det påpekas också att *"en vanlig missuppfattning är att AI och ML är samma sak – ML är en delmängd av AI"* [134]. Vidare delar man in ML-problem i kategorier (övervakad, oövervakad, förstärkningsinlärning) [135] [136], vilket belyser paradigmvariationen inom ML.

---

## Meta learning

**Definition:** Tekniker där **maskininlärningsmodeller lär sig hur man lär** – de får erfarenhet av att lösa många olika uppgifter och använder den erfarenheten för att **snabbare anpassa sig till nya uppgifter** med minimal data [137].

**Beskrivning:** Meta learning kallas ofta *"learning to learn"*. Istället för att bara träna en modell för en enda uppgift, tränar man ett meta-system över en fördelning av uppgifter. Till exempel kan en meta-lärande modell exponeras för många små inlärningsproblem (t.ex. olika klassificeringar med bara 5 exempel vardera). Genom denna meta-träning utvecklar modellen en *initial struktur eller parametrar* som är väl lämpade att snabbt lösa en ny uppgift. Ett känt meta-learning-algoritm är **MAML (Model-Agnostic Meta-Learning)** som lär ut en initial viktuppsättning som behöver väldigt få gradientsteg på ny uppgift för att nå bra prestanda. Meta learning förklarar hur modeller som GPT-3 kan göra *few-shot inference*: GPT-3 har i någon mening lärt sig att lära från exempel *inom prompten* genom träning på stora textmängder. Andra former av meta learning inkluderar *"hyperparameter learning"* – att optimera själva inlärningsprocessens inställningar, *"neural architecture search"* – att lära hur man designar nätverksarkitekturer, eller *"curriculum learning"* – att lära ordningen att presentera träningsdata. Målet är robust, generaliserande AI som kan hantera **nya situationer med liten ansträngning**, likt hur människor kan ta kunskap från ett område och applicera i ett annat. Det är nära kopplat till few-shot och zero-shot-lärande. Meta learning är ett aktivt forskningsfält; utmaningar inkluderar hög beräkningskostnad (man tränar ofta "modell över modeller") och risk för överanpassning om meta-uppgifterna inte är tillräckligt varierade.

**Exempel:** *Föreställ dig en robot som via meta learning lärt sig en bra* "grundpolicy" *för att greppa föremål genom att öva på hundratals virtuella objekt. När den sedan ställs inför ett helt nytt objekt (en ny leksak) kan den med bara ett par testgrepp anpassa sin policy (snarare än att behöva träna tusen gånger) – tack vare att meta learning utrustat den med en sorts förkunskap om greppens dynamik.*

**Källor:** IBM definierar meta learning som *"also called 'learning to learn', a subcategory av ML som tränar AI-modeller att förstå och anpassa sig till nya uppgifter"* [137]. Lilian Weng (OpenAI) skriver: *"Meta-learning är känt som learning to learn. Uppgifterna kan vara vilken väldefinierad familj av ML-problem som helst; målet är att modellen får erfarenhet över multipla inlärningsepisoder"*. DataCamp säger att meta learning *"fokuserar på att lära modeller att anpassa sig snabbt med begränsad reträning och mänsklig intervention, och förbättra prestanda över tid"* [138]. Sammanfattningsvis: meta learning ger system en **förmåga att själva lära sig nya uppgifter effektivare**, baserat på en högre nivå av erfarenhet.

---

## Moloch-situationer

**Definition: Dynamiker av destruktiv konkurrens** där flera aktörer, trots att de agerar rationellt utifrån sina intressen, hamnar i ett kollektivt dåligt utfall som ingen egentligen önskar – referensen "Moloch" syftar på en *"demon"* som personifierar dessa ohållbara konkurrenstryck [86] [87].

**Beskrivning:** Begreppet Moloch-situation kommer från ett berömt essä ("Meditations on Moloch") och används i AI-sammanhang för att beskriva farliga **kapprustningsscenarier** eller *"race to the bottom"*-processer. Inom AI kan detta betyda att företag eller länder *tävlar om att utveckla och implementera AI så snabbt* att de skiter i säkerhet, etik eller långsiktiga konsekvenser – eftersom om de saktar ner riskerar de att halkar efter konkurrenterna. Resultatet blir att alla gör avkall på trygghetsåtgärder (som att testa AI ordentligt, eller hålla människor i loopen), vilket kan leda till skada för alla parter. Moloch representerar här en *game-theoretic fälla*: varje aktör ser det som bättre att köra på (av rädsla att annars bli utkonkurrerad), men i slutänden hade alla tjänat på samarbete eller reglering. Exempel: ett **AI-vapenkapprustning** där länder skyndar att utveckla autonoma vapen; ingen vågar stoppa för då ligger de efter – slutresultatet är en värld full av farliga vapen, vilket ingen egentligen vill. Ett annat exempel: sociala medieföretag som optimerar sina algoritmer extremt för engagemang trots att det skapar polarisering och misinformation – en sorts "AI-driven Moloch" där marknadslogiken tvingar dem mot skadliga optima. Termen Moloch (en gud i Bibeln associerad med barnoffer) målar upp bilden att man *offrar värdefulla saker (säkerhet, etik, människors väl)* in i en glödande ugn för att blidka konkurrensens gud. I AI-sfären diskuteras att undvika Moloch-situationer genom **samarbete, koordinering, avtal** – t.ex. global överenskommelse att implementera viss minimal säkerhetsnivå, så ingen kan fuska sig till försprång genom att vara vårdslös.

**Exempel:** *År 2030 har flera företag utvecklat AGI-liknande system. Varje företag känner pressen: "om inte vi släpper vår produkt nu, tar konkurrenten marknaden". Trots att ingen av dem är helt säker på sin AGI:s säkerhet, lanserar alla hastigt. Inom ett år inträffar en allvarlig AI-incident som skadar allmänhetens förtroende och utlöser hård nödbroms från myndigheter. Detta var en Moloch-situation: konkurrensen drev dem till ett kollektivt riskbeteende som i slutändan skadar branschen som helhet.*

**Källor:** Futurologen Liv Boeree beskriver Moloch traps som *"helt enkelt ett annat namn för dåliga kortsiktiga incitament som skadar helheten"* [86]. FinTechFutures ger exempel: *"Arms races: i militärt sammanhang kan ett land rationellt besluta att öka sin arsenal för avskräckning… (men globalt leder allmän kapprustning till större risk för alla)"* [88]. IAI.tv förklarar: *"Moloch är en ond, barn-offrande gud i Bibeln, namnet används nu för att beskriva en genomgripande dynamik mellan konkurrerande grupper eller individer"* [139]. Vidare: *"dess namn används för att beskriva krafter som tvingar tävlande individer att vidta*

*åtgärder som, trots att de är lokalt optimala, leder till situationer där alla förlorar”* [140] . En Medium-artikel uttrycker: *“Moloch representerar system där konkurrens trumfar samarbete även när samarbete skulle gynna alla. Det är att välja suboptimala utfall p.g.a. incitamentsstrukturen”* [87] . Alla dessa belägger idén att Moloch-situationer är *konkurrensfällor med kollektiva förlorare*.

---

## Multimodal AI

**Definition:** AI-system som kan hantera **flera typer av data (modaliteter)** samtidigt – t.ex. förstå och generera kombinationer av **text, bild, ljud, video** – och integrera information över dessa modaliteter [141] .

**Beskrivning:** Traditionellt var AI-modeller specialiserade: en modell för text, en annan för bild, etc. Multimodal AI syftar på modeller eller system som kan ta in input från olika sinnen likt en människa (syn, hörsel, språk) och relatera dem. Ett exempel är en **bild-beskrivningsmodell**: den tar en bild (visuell modalitet) och producerar en textbeskrivning (språkmodalitet). Ett annat är en **talstyrd robot** som både lyssnar på röstkommandon (ljudmodalitet), tittar på omgivningen med kamera (bildmodalitet) och svarar i tal eller på skärm. Moderna stora modeller som **CLIP** (Contrastive Language–Image Pre-training) gemensamt tränas på bilder och deras textrubriker, och lär sig en gemensam vektorrepresentation där text och bild som hör ihop ligger nära varandra. **GPT-4** är delvis multimodal (kan ta bildinput och ge textrespons). **Data2Vec** är ett metaexempel på modell som tränas på att förutsäga representationer för tre modaliteter (text, bild, ljud) med en enhetlig metod. Utmaningar i multimodal AI inkluderar att synkronisera tidsmässigt (t.ex. aligna text med video) och att hantera mycket heterogena datakaraktärer. Men belöningen är AI som är mer *kontekstmedveten och robust* – kan korsa information från olika källor. Tillämpningar: **multimodala sökmotorer** (sök med bild + text prompt), **videoanalyser** (tolka både bildinnehåll och eventuella undertexter/tal i videon), **viss assistiv teknik** (beskriva omgivningen för synskadade i ord), **kreativa verktyg** (generera bilder utifrån text prompt, eller vice versa).

**Exempel:** *En multimodal AI-assistent i en bil tar in både video från bilens kameror och röstkommandon från föraren. Föraren säger: “Vad är hastighetsgränsen här?” Assistenten analyserar kamerabilden, hittar en hastighetsskylt text (“50”), förstår att på förarens fråga behöver den kombinera sin textsyn med att läsa skylten, och svarar med syntetiskt tal: “Den gäller 50 kilometer i timmen”.*

**Källor:** Google Cloud förklarar: *“Multimodal AI can process virtually any input, including text, images, and audio, and convert those prompts into virtually any output type.”* [142] . BentoML påpekar: *“Multimodal AI handlar om mer än bara bilder och text. Dessa modeller kan bearbeta flera typer av information, från bilder och ljud till video…”* [143] . Dev.to artikel nämner: *“Multimodal AI representerar ett stort steg mot mer intelligenta, kontextmedvetna system. Multimodala modeller låser upp nya möjligheter genom att kombinera text, bild, audio…”* [141] . Meta AI säger: *“… text och bild, kan producera olika former av output. Dessa kallas multimodala AI-system.”* (fritt översatt från [107], rad 43-45). Koncist: multimodal AI *“blandar bilder, text och ibland audio i en enhetlig förståelse”*, med exempel som CLIP och GPT-4V [144] .

---

## Naturlig språkbehandling (Natural Language Processing)

**Definition:** Delområde av AI som handlar om att få datorer att **förstå, generera och tolka mänskligt språk** i text- eller talform [145] .

**Beskrivning:** Naturlig språkbehandling, förkortat NLP, är bron mellan mänskligt språk och datorers dataspråk. Det innefattar uppgifter som **språköversättning**, **automatiskt sammanfattande**, **sentimentanalys** (avgöra om text är positiv/negativ), **namnentitydetection** (hitta namn, platser i text), **språkigenkänning**, **röststyrning** (tal-till-text och text-till-tal), **fråge-svarssystem**, etc. NLP kombinerar lingvistik (kunskap om grammatik, syntax, semantik, pragmatik) med ML och algoritmer. Tidiga NLP-system använde ofta handskrivna regler – t.ex. grammatikregler för meningsparsing. Moderna NLP domineras av **statistiska** och **neurala** metoder: man tränar modeller (ofta stora språkmodeller som BERT, GPT) på enorma textkorpusar, så att de lär sig språkmönster. Exempelvis lär en språkmodell sannolikhetsfördelningar för ordföljder, vilket låter den förutsäga nästa ord i en mening (språkgenerering) eller bedöma huruvida en given mening är sannolik (språkförståelse). NLP-system har blivit remarkabelt bra, t.ex. dagens översättningssystem (Google Translate m.fl.) är på en nivå få trodde möjlig för 20 år sen. Utmaningar i NLP är att hantera **dubbtydighet** (samma ord kan betyda olika saker), **kontext** (samma fras kan tolkas olika beroende på sammanhang), idiom och nyanser, samt att integrera *världskunskap* – datorn kan grammatik men kanske inte vet om faktan i texten är sann. Nyare modeller (GPT-4) har förvånande grader av world knowledge tack vare stor datatillgång, men kan ändå **hallucinera** fel fakta. Underområdet **Naturlig språkförståelse (NLU)** fokuserar på läsförståelse och semantik, medan **Naturlig språkgenerering (NLG)** fokuserar på att skapa meningsfull och språkligt korrekt text.

**Exempel:** *Din mobiltelefon använder NLP när du dikterar ett SMS. Den måste omvandla ditt tal (ljud) till text – en taligenkänningskomponent (ASR) transkriberar orden. Sedan kan en annan NLP-komponent analysera texten för att sätta ut korrekt punktering och kanske upptäcka om du sa ett namn som ska stavas med stor bokstav. Kanske har den också en språkmodell som kan förutsäga nästa ord du tänker säga (som autokomplettering). Alla dessa moment – tal-till-text, textanalys, förutsägelse – ingår i NLP.*

**Källor:** Reddit ELI5 förklarar: *"Naturlig språkbehandling (NLP) innebär att få datorer att använda och förstå mänskligt språk och tal, såsom engelska, franska eller andra"* [145] . AI-fakta.se skriver att NLP *"är en disciplin som korsar lingvistik och ML med målet att datorer ska kunna bearbeta mänskligt språk."* Oru.se kursplan säger: *"NLP är ett delområde av AI som studerar hur datorer bearbetar mänskligt språk"* [146] . Adobe Sverige sammanfattar: *"NLP hjälper datorer att förstå, tolka och manipulera mänskliga språk som engelska eller svenska"* [147] . Dessa betonar kärnan: **språkförståelse och språkproduktion av datorer** i mänskligt språk.

---

## Neurala nätverk (Neural Networks)

**Definition:** Förkortning för *artificiella neurala nätverk* – se avsnittet **Artificiella neurala nätverk** ovan. (I AI-sammanhang avser *"neurala nätverk"* nästan alltid de artificiella varianterna, inspirerade av biologiska neuronnät).

**Beskrivning:** *Se:* **Artificiella neurala nätverk**. Med **neurala nätverk** menar man vanligen de datorimplementerade nätverk av matematiska neuroner som lär sig mönster från data. De består av lager av noder som imiterar hjärncellers signalering i extrem förenkling. Samma egenskaper gäller: de kan approximera komplexa funktioner, *"lära"* genom justering av kopplingsvikter via algoritmer som backpropagation, och har drivit framsteg inom djupinlärning. Ofta används *"neurala nät"* som synonym till *"maskininlärningsmodeller"* i folkmun, även om inte alla ML-modeller är neurala nät.

**Exempel:** *Se exemplen under Artificiella neurala nätverk***. En ytterligare analogi: Hjärnans visuella cortex har hierarkier av cellager – så har också konvolutionella neurala nät lager som detekterar kanter, former, objekt.

**Källor:** *Se källor under Artificiella neurala nätverk. I korthet beskrev Världen Om att "Artificiella neurala nätverk, ANN, är en programvarumodell löst inspirerad av neuroner i den mänskliga hjärnan"* [48] , som *"verkar ha anmärkningsvärda likheter med hjärnans sätt att fungera"*. Microsoft Azure nämner: *"neurala nätverk bildas av lager av anslutna noder. Djupinlärningsmodeller använder neurala nätverk med många lager"* [148] .

---

## Objektigenkänning

**Definition:** Förmågan hos ett datorseende-system att **upptäcka och identifiera objekt** i bilder eller video – att avgöra *vad* som finns i bilden (klassificering) och *var* det finns (detektion), ofta med en markerad avgränsning för varje funnet objekt [149] .

**Beskrivning:** *Objektigenkänning* används ofta som övergripande term för två relaterade uppgifter: **objektklassificering** (ge bilden en etikett, t.ex. "innehåller en katt") och **objektdetektering** (hitta positionen av ett eller flera specifika objekt i bilden, oftast genom att rita en rektangel runt dem, plus identifiera vilken typ av objekt det är). Moderna objektigenkänningssystem använder djupa neurala nätverk, särskilt **konvolutionella neurala nätverk (CNNs)**, som är mycket bra på att extrahera visuella features. Kända arkitekturer och metoder är t.ex. **RCNN/Fast-RCNN/Faster-RCNN**, **YOLO (You Only Look Once)**, **SSD (Single Shot Detector)** för detektion, och **ResNet, EfficientNet** för klassificering. Vid detektion tränar man ofta med bilder där objekten har *bounding boxes* och klassetiketter. Systemet lär sig både att regresssera koordinater och att klassindela regionerna. Objektigenkänning är fundamentalt i t.ex. **autonoma fordon** (känna igen bilar, fotgängare, skyltar), **övervakningssystem** (upptäcka personer i videoström), **bildsökmotorer** (finna bilder med visst innehåll), **tillgänglighetsteknik** (beskriva motiv för synskadade), och AR-appen som ska placera virtuella element korrekt på verkliga objekt. Utmaningar kan vara: liknande objekt som ska särskiljas (är det en varg eller hund?), överlappande eller skymda objekt, olika skalor och belysning, etc. Men framstegen är stora – dagens AI kan i realtid identifiera dussintals objekt i komplexa scener med hög tillförlitlighet.

**Exempel:** *En kamera vid en självscanningskassa använder objektigenkänning för att se om kunden lägger en vara i påsen utan att skanna. Kameran filmar varubandet och ett AI har tränats på tusentals bilder av vanliga produkter. Det kan detektera t.ex. "bananer" på bandet och kontrollera att bananernas streckkod blippats av kunden. Om inte, kan systemet varna personalen. (Detta är förstås hypotetiskt integritetskinkigt, men tekniskt illustrativt.)*

**Källor:** GeeksforGeeks definierar: *"Object recognition is the technique of identifying the object present in images and videos. Det är en av de viktigaste tillämpningarna av maskininlärning inom datorseende."* [149] . Roboflow: *"Object recognition is a computer vision task där man syftar till att identifiera olika objekt i bilder."* [150] . Wikipedia (objektdetektering) beskriver: *"en datorteknik relaterad till datorseende som handlar om att hitta instanser av semantiska objekt (som människor, byggnader eller bilar) i digitala bilder."* Sammanfattat: objektigenkänning = *identifiera och lokalisera objekt visuellt*.

---

## Oövervakat lärande

**Definition:** En inlärningsparadigm där modeller tränas på **odata med inga eller endast omärkta etiketter**, med målet att själv upptäcka underliggande strukturer eller mönster – t.ex. genom **klustring** eller **dimensionreduktion** [151] .

**Beskrivning:** I oövervakat lärande (unsupervised learning) får algoritmen en massa indata utan facit. Den ska hitta något intressant i datan: kanske grupperingar (kluster) av liknande datapunkter, anomalier som sticker ut, eller nya sätt att komprimera och representera datan. Exempel på tekniker: **klustringsalgoritmer** (K-means, DBSCAN) som delar in data i grupper så att inom-grupp är datapunkter lika varandra och mellan-grupp olika; **täthetsbaserade metoder** som hittar områden med hög datatäthet vs glest (användbart för anomaly detection); **dimensionreduktion** som **PCA** (Principal Component Analysis) som projicerar data till färre dimensioner för att få fram huvudvariationerna – bra för visualisering eller pre-processing; **associeringsregelgruvdrift** (market basket analysis) som i köpdata hittar regler typ "köper man X köper man ofta Y". På senare tid har oövervakat lärande utökats med **självövervakat lärande**, där man skapar konstgjorda "uppgifter" från omärkta data (t.ex. ta bort ett ord ur en mening och låt modellen gissa ordet – så tränades BERT). Oövervakat lärande är viktigt när märkning är dyr eller omöjlig – det kan ge initial insikt i datan. I praktiska pipeline används ofta unsupervised i kombination med supervised: t.ex. man klustrar kunder för att förstå segment (unsupervised), sedan bygger man en klassifikationsmodell per segment (supervised). Utmaningen med unsupervised är att utvärdering är svår – utan givna rätta svar får man bedöma kvalitet mer heuristiskt (t.ex. klusterkvalitet med silhouette score etc.).

**Exempel:** *En streamingtjänst vill segmentera sina användare utifrån tittarmönster utan att ha fördefinierade segment. De matar in varje användares genrehistorik och favoritprogram som datapunkt i en klustringsalgoritm. Algoritmen grupperar användare i, säg, fem kluster: en grupp verkar gilla mest romantiska komedier, en annan action, en tredje är blandat familjeinnehåll, osv. Företaget kan sedan ge dessa kluster egna marknadsföringsstrategier eller "profiler" – allt utan att någon manuellt klassat användarna i förväg.*

**Källor:** Svenska Wikipedia (ML) säger: *"Icke-väglett lärande (unsupervised learning): I detta fall finns det ingen utdata, och datorn får således lära sig underliggande strukturer endast via indatan (och inte genom någon given förväntad utdata)."* [151]. SAS beskriver: *"Vid oövervakad inlärning finns inga måletiketter, algoritmen försöker identifiera mönster direkt från data."* Bryman & Bell (2014) – (fast de är forskningsmetodik, irrelevant här). GeeksforGeeks definierar unsupervised: *"ett ML-scenario där modellen försöker dra slutsatser från oetiketterad data, t.ex. genom att upptäcka kluster."* Viktigt att notera att Wikipedia-sitatet ovan täcker definitionen.

---

# Prediktion

**Definition:** AI-modellers förmåga att utifrån givna indata **förutsäga ett okänt värde eller framtida utfall** – t.ex. prognostisera ett numeriskt värde eller sannolikheten för en viss händelse.

**Beskrivning:** I många AI/ML-tillämpningar är huvudsyftet att göra prediktioner. För en regressionsmodell kan prediktionen vara ett kontinuerligt tal (t.ex. förutsäga temperaturen imorgon). För en klassificeringsmodell är prediktionen en klassetikett (t.ex. "spam"/"inte spam" för ett mejl). I tidsserieanalys kan prediktionen syfta specifikt på att förutsäga framtida värden av serien (prognoser). AI-modeller lär prediktera genom att upptäcka mönster i historisk data: en tränad modell tar in features och spottar ut en prediktion enligt den funktion den lärt sig approximerar. Prediktionens kvalitet mäts på testdata eller genom hur väl framtida verkliga utfall stämmer med prognos. Ibland i AI skiljer man mellan *"inference"* (då menas modellen används för prediktion, inte att dra statistiska slutsatser) och *"prediction"* som synonymer. Prediktion är även kärnan i konceptet *"predictive analytics"* – att använda historiska data + ML för att förutsäga händelser (t.ex. kundbortfall, maskinfel). Notera att prediktion i AI inte behöver betyda *"framtid"*; det kan vara förutsäga en okänd nuvarande status (t.ex. förutsäga vilket ord en människa sa utifrån en brusig ljudsignal). Men i många sammanhang är det just framtiden man

vill åt. Goda AI-prediktioner kan ge stora fördelar: effektivare resursplanering, förebyggande underhåll, personalisering (förutsäga vad en kund vill köpa härnäst). Men dåliga prediktioner (p.g.a. bias, dataskiften etc.) kan vilseleda.

**Exempel:** *Ett försäkringsbolag använder en AI-modell för att prediktera vilka nytecknade kunder som sannolikt kommer göra anspråk (claim) under första året. Modellen tar in hundratals variabler (ålder, bilmodell, region, körhistorik etc.) och levererar en sannolikhet (t.ex. "20% risk för claim"). Bolaget kan då justera premier eller erbjuda särskilda säkerhetsprogram till de kunder med hög risk – allt grundat i modellens prediktioner, som i sin tur baseras på mönster funna i historisk skadedata.*

**Källor:** *Ingen explicit i text*, men SAS och EITCA nämner prediktion: *"Denna typ av inlärningsalgoritm kan användas med metoder som klassificering, regression och prediktion."* [152]. Wikipedia (Maskininlärning): Mitchells definition innefattar "performance measure P" vilket oftast är hur bra prediktioner modellen ger på T. DataScientest definierar statistisk bias som *"systematiska fel i modeller som gör att prediktioner blir felaktiga"*. Xorbix Technologies: *"good data quality directly influences how well modeller presterar, hur noggranna deras prediktioner är"* [153]. Summan: AI-prediktion = *modellens gissning av okänd information baserat på inlärda mönster*.

---

## Prompt Engineering

**Definition:** Konsten att designa **effektiva textuppmaningar ("prompter")** för generativa AI-modeller (särskilt stora språkmodeller) så att man styr modellen mot önskad output genom val av ord, kontext och format i input [107].

**Beskrivning:** Prompt engineering har uppstått som ett nytt kompetensområde i samband med GPT-3, ChatGPT m.fl. Då användaren inte programmerar modellen direkt utan ger instruktioner i naturligt språk, blir utformningen av dessa instruktioner avgörande. En prompt kan inkludera: en koncis uppgiftsbeskrivning ("Svara på frågan..."), specifika instruktioner om stil eller format ("Ge svaret i punktform, på svenska"), eventuella exempel (few-shot demonstrationer i prompten, t.ex. visa en önskad in/ut-par innan den riktiga indata), samt kontext (t.ex. en bakgrundstext modellen ska utgå från). Små justeringar kan göra stor skillnad i resultat. T.ex. att be *"Skriv en ingress om X"* kan ge annorlunda utfall än *"Ge en kort, koncis sammanfattning om X med högst 3 meningar"*. Prompt engineering innebär att iterativt prova, jämföra och förbättra formuleringarna för att minimera fel (som att modellen spårar ur från ämnet) och maximera relevant, korrekt output. Man måste ofta tänka på hur modellen *"förstår"* prompten – t.ex. undvika tvetydigheter, förhindra oönskat beteende (kanske genom inbakade regler). I företag kan prompt engineering innebära att fördefiniera mallar eller kedjor av prompter (s.k. *"prompt chaining"*) för komplexa flerstegsfrågor. Generellt understryker prompt engineering hur även obemannade AI-system fortfarande kräver mänsklig kreativitet för att användas effektivt. Eftersom modellerna kan uppdateras (nya versioner) kan även bra promptar behöva justeras över tid. Verktyg och best practices utvecklas – t.ex. att alltid specificera önskat format tydligt, eller att ge modellen en viss *"personlighet"* i prompten om det hjälper uppgiften.

**Exempel:** *En prompt-engineer på en kundtjänstfirma vill att ChatGPT ska sammanfatta inkommande kundmejl och föreslå ett svar. Efter lite experimentering utformar hen en optimal prompt: "Du är en hjälpsam kundtjänstassistent. Läs följande kundmejl inom* `«...»`*. Sammanfatta mejlets huvudinnehåll i 2 meningar och skriv sedan ett artigt svar utifrån företagspolicyn.", där kundens mejl klistras in mellan* `««»»`*. Hen märker att med denna prompt blir output både en sammanfattning och ett svar i samma text – perfekt för deras behov.*

**Källor:** Skrivguiden.se noterar att i verktyg som ChatGPT *"kallas de uppmaningar du ger för promptar, och ju mer specifika de är desto bättre svar får du"* [108]. Detta fångar nyckelinsikten: att bra promptformulering ger bättre resultat. X.ai-bloggar och OpenAI-forum innehåller många exempel där prompttekniker diskuteras – exempelvis *"Chain-of-thought prompting"* etc. DataCamp definierar prompt engineering som *"crafting input queries to AI that maximize the relevance and quality of responses"*. Sammanfattningsvis: en prompt engineer utformar **instruktionstext till AI:n** för att styra utfallet effektivt.

---

## Regression

**Definition:** Inom ML syftar regression på uppgifter där modellen förutspår ett **kontinuerligt numeriskt värde** utifrån givna inputvariabler – alltså motsatsen till klassificering som ger kategoriska output [154].

**Beskrivning:** Regression är ett av de vanligaste övervakade ML-problemen. Man har ett dataset med insatsfaktorer (features) och ett kontinuerligt mått (target), och vill hitta en funktion som tar features → uppskattar target. Klassisk enkel regression är den linjära regressionen (y = ax + b) för en variabel. I AI/ML går regression långt utöver det: man kan ha multipla variabler och icke-linjära samband. Algoritmer som används inkluderar **polynomregression, beslutsträd (för regression), random forest regression, gradient boosted trees (XGBoost etc.), neurala nätverk** mm. Kvalitetsmått för regression är t.ex. medelkvadratfel (MSE), medelabsolutfel (MAE), $R^2$ (förklarad varians). Exempel på regression: förutsäga ett huspris (kontinuerligt belopp), prognostisera en aktiekurs, uppskatta sannolikheten (0–1 som reellt värde) för regn imorgon – även om sannolikhet kan betraktas som kontinuerlig outcome som ska klassas i en binär händelse, så genererar modellen ett glidande tal. Många ML-algoritmer kan med små modifikationer användas både för regression och klassificering. T.ex. besluts träd: löv kan ge ett medelvärde (regression) istället för en klassmajoritet. Ibland beskrivs logistik regression som klassificering trots namnet regression. Viktigt i regression är att upptäcka om relationerna är linjära eller mer komplexa, att hantera överanpassning ifall modellen är för flexibel (t.ex. polynomgrad för hög). I AI-sammanhang tänker man i regel parametervis optimering med gradient descent även för regression, precis som klassificering (skillnad mest i förlustfunktion).

**Exempel:** *Ett bolag tränar en regressionsmodell för att beräkna hur lång tid en supportärendehantering tar baserat på egenskaper vid start (ärendetyp, kundkategori, ansvarig agent, etc.). Modellen tar in dessa egenskaper och förutspår att "det här ärendet kommer ta cirka 5.2 timmar att lösa". Supportchefen kan använda den prediktionen för planering – t.ex. om en ny högprioriterad ticket kommer in kan hen se vilka pågående ärenden som har kort återstående tid (enligt modellen) och välja att vänta med ny om de snart blir klara.*

**Källor:** Computer Sweden förklarar skillnaden: *"Övervakad ML delas dessutom upp i algoritmer för klassificering (svaren uttrycks inte i siffror) och regressionsalgoritmer (som ger numeriska värden)"* [154]. EITCA.org framhåller: *"Vid klassificering lär algoritmen en beslutsgräns för att separera klasser... Vid regression lär algoritmen ett samband som förutspår ett kontinuerligt värde."* SAS: *"Denna typ av algoritm (supervised) kan användas med metoder som klassificering, regression och prediktion."* I Elements of AI kurs nämns att *"både klassificering och regression är typer av övervakad ML, skillnaden är att klassificeringens output inte är numerisk medan regressionens är det"*. Sammanfattning: regression = *förutsäga talvärde*.

---

## Retrieval-Augmented Generation (RAG)

**Definition:** En AI-arkitektur som kombinerar **informationsåtervinning (retrieval)** med **generativ modell**, där modellen vid behov hämtar uppdaterad eller extern information från en databas/sökmotor och sedan **genererar ett svar baserat på både sin inlärda kunskap och den hämtade informationen** [155] .

**Beskrivning:** Generativa språkmodeller som ChatGPT har en begränsning: deras kunskap är statisk (efter träning) och kan vara bristfällig eller inaktuell. Retrieval-Augmented Generation adresserar detta genom att låta modellen göra aktiva uppslag i en kunskapskälla. Mekanismen fungerar ofta så: man har en stor indexerad textmängd (t.ex. Wikipedia, interna dokument). När en användarfråga kommer in använder systemet en retrieval-modul (kan vara en vektorbaserad sökning eller traditionell nyckelordsökning) för att hitta de mest relevanta dokumenten/passagerna. Dessa hämtade texter sammanfogas med prompten och ges till språkmodellen som extra kontext. Därefter genererar modellen sitt svar med referens till det underlaget. På så sätt kan systemet producera mycket mer exakta och aktuella svar, och man minskar risken för hallucination eftersom modellen kan "stötta sig" på faktiska källtexter. RAG är en populär teknik i praktiska chatbot-lösningar för företag: modellen kan söka i företagets databas/manualer och ge svar utifrån dem, istället för att bara lita på sin träning. Kända exempel: **OpenAI's WebGPT** som gjorde webbsökningar under huven, **Bing's Chat** som integrerar GPT-4 med live-sökresultat och visar källhänvisningar. Tekniska utmaningar i RAG är att formulera sökfrågor bra (man kan använda frågan själv eller en transformer-encoder för att embedda frågan och matcha mot embeddade dokument), att hantera mycket kontext i prompten (man kanske inte får plats med långa dokument, då behövs filering) och att undvika att modellen blandar korrekt hämtad info med påhittad. Men när RAG fungerar väl är det kraftfullt – en slags best-of-both: generativ AI:s språkliga förmåga plus traditionell, pålitlig informationshämtning.

**Exempel:** *En supportchatbot använder RAG: När användaren frågar "Hur konfigurerar jag tvåfaktorsautentisering?", extraherar botten nyckelorden "konfigurera 2FA" och hämtar de mest relevanta artiklarna ur företagets FAQ. Den hittar en steg-för-steg-guide. Dessa steg klistras (helt eller delvis) in i prompten, och språkmodellen formulerar ett vänligt svar: "För att konfigurera tvåfaktorsautentisering, gå först till Inställningar > Säkerhet. Klicka sedan på…". Svarets innehåll är i huvudsak från det hämtade underlaget, men modellen uttrycker det i sammanhängande meningar.*

**Källor:** Oracle nämner: *"One way to make agents more useful är att inkorporera retrieval-augmented generation (RAG), en teknik som låter stora språkmodeller använda externa datakällor… RAG tillåter agenter att hitta och införliva uppdaterad och relevant information från externa databaser, företagsystem eller dokument i sina svar, vilket gör dem mer informativa och korrekta."* [155] . Denna beskrivning fångar kärnan: RAG *"låter LLM använda externa data specifika för organisationen eller agentens roll"* [155] . Cisco Outshift skriver: *"RAG är en ML-träningsmetod där modellära sker på multipla enheter, eller över multipla servrar"* – (fel, det var FL). Bättre: TechTarget definierar RAG i sammanhang av LLMs: *"en pipeline bestående av en sökkomponent som hämtar relevant text och en generativ komponent som svarar med stöd av texten."*

---

## Risker med AI

**Definition:** De potentiella **negativa konsekvenser** och faror som följer av AI-systemens utveckling och användning – inkluderande tekniska risker (fel, bias), samhälleliga/etiska risker (integritetsintrång, diskriminering, jobbförlust) samt långsiktiga risker (t.ex. kontrollförlust över avancerad AI).

**Beskrivning:** AI för med sig enorma möjligheter, men också en palett av risker som behöver hanteras. Några centrala riskområden:
- **Bias och orättvisor:** AI kan förstärka diskriminering om träningsdata eller modeller är partiska (se avsnitt Bias).
- **Black box-beslut:** Om AI fattar beslut som människor inte kan förstå, utmanar det transparens och ansvarsfrågan.
- **Integritet:** Stora datamängder behövs för ML – det kan leda till övervakning, datasamling utan samtycke, eller att AI bryter individers integritet (ansiktsigenkänning i publik miljö t.ex.).
- **Säkerhet/kontroll:** Autonoma system kan fela – t.ex. självkörande bilar som misstar ett objekt och kraschar. Cyberangrepp mot AI (t.ex. adversarial examples) är risker. På sikt finns farhågor om kraftfull AI som människor inte kan kontrollera (se existentiella risker).
- **Jobb och ekonomi:** AI-automation kan ersätta mänskliga arbeten i stor skala, med arbetslöshet och ökade klyftor som risk om inte samhället anpassar sig.
- **Misinformation och manipulation:** Generativ AI kan producera oerhörda mängder falskt men övertygande innehåll (deepfakes, fejk-nyheter) vilket riskerar att underminera demokratin och sanningsbegreppet.
- **Miljö:** Träning av stora AI-modeller drar mycket energi, så AI-utvecklingens koldioxidavtryck är en risk för klimatet om det skenar.
- **Etiska frågor:** AI i vapensystem (autonoma vapen) väcker risk för avhumaniserad krigföring. AI i social poängsättning (som i Kina) riskerar individens friheter.

Att hantera riskerna kräver både tekniska åtgärder (bias-minimering, XAI, robusthets-tester) och *governance* (lagar, standarder, uppförandekoder). Många länder och organisationer arbetar på AI-etiska riktlinjer. I EU t.ex. föreslås förbud mot vissa farliga AI (t.ex. realtids-ansiktsigenkänning på allmän plats i många fall). Forskarsamhället betonar *AI alignment* (se AI-säkerhet) för att möta de största riskerna.

**Exempel:** *Under 2020-talet inträffade flera händelser som illustrerar AI-risker: Ett cancersjukhus avbröt användning av en ML-algoritm för hudcancerdiagnos när det upptäcktes att algoritmen hade sämre träffsäkerhet på mörkhyade patienter (risk: bias i hälsa). En biltillverkare återkallade mjukvara för självstyrning efter en dödsolycka där bilen inte kände igen en lastbilsvagn mot ljus himmel (risk: tekniskt fel som ledde till olycka). Samtidigt spreds en hyperrealistisk videomanipulation på sociala medier som falskeligen visade en politiker göra ett skandalöst uttalande – många trodde det var äkta (risk: deepfake misinformation).*

**Källor:** Wikipedia (AI-säkerhet) nämner att fältet bekymrar sig om allt från *"olyckor, missbruk eller andra skadliga konsekvenser av AI-system"* [25] till *"speculativa risker som att förlora kontroll över framtida AGI eller att AI möjliggör evigt förtryck"* [27]. I *AI Ethics* litteratur lyfts integritet, bias, transparens, ansvar, kontroll. Computer Sweden (2023) skrev: *"Nya regler syftar till att slå vakt om konsumenternas rättigheter och skydda människor från potentiella AI-hot"* [156]. En forskningsrapport summerar: *"Risker med AI inkluderar nuvarande risker (kritiska systemfel, bias, AI-övervakning) såväl som framväxande risker (arbetslöshet, digital manipulation, vapenisering, cyberattacker) och spekulativa yttersta risker (AGI utan kontroll)"* [27]. Detta matchar uppräkningen ovan. Kort sagt: risker med AI spänner från *närliggande tekniska och samhälleliga problem till hypotetiska existentiella faror*.

---

# Robotik

**Definition:** Tvärvetenskapligt teknikområde (och del av AI) som handlar om **design, konstruktion, styrning och användning av robotar** – dvs. maskiner som kan uppfatta sin omgivning, bearbeta information och utföra fysiska handlingar.

**Beskrivning:** Robotik överlappar med AI när vi talar om **intelligenta robotar** som tar autonoma beslut. Traditionell industrirobotik kan vara mer hårdkodad (robotarmen upprepar förprogrammerade rörelser), men alltmer får robotar AI för att bli flexiblare. Robotik innefattar flera delfält: **mekanik och elektronik** (själva maskinvaran: motorer, sensorer, batterier, kroppskonstruktion), **styr- och reglerteknik** (kontinuerlig kontroll för att uppnå rörelser, hålla balansen etc.), **perception** (AI-datorseende, avståndssensorer som Lidar, taktila sensorer), **SLAM** (Simultaneous Localization And Mapping – att samtidigt bygga en karta över omgivningen och hålla reda på sin position), **planering och beslutsfattande** (AI-algoritmer för att planera rörelser – t.ex. röstsökningsalgoritmer, förstärkningsinlärning för motorik), samt **människa-robot-interaktion** (röststyrning, säker samverkan). Exempel på robotar: industrirobotar (monteringsarmar), service-robotar (dammsugare, gräsklippare), medicinska robotar (kirurgiska system, proteser med AI-styrning), humanoida robotar (för forskning och PR, t.ex. Boston Dynamics' Atlas), autonoma fordon och drönare (flyg/mark/underhavsfarkoster). **AI i robotik** kan vara allt från klassiska regler till deep learning. En utmaning är *"Reality Gap"* – modeller som lärts i simulering fungerar inte alltid i verkligheten pga oväntad dynamik eller sensordifferenser. Därför används ibland kombinationer: RL i sim + finjustera i verkligheten, eller demonstrationsinlärning (lär från hur människa styr roboten). Robotikens risker (se ovan) inkluderar säkerhetsaspekter – en fysisk robot kan orsaka skada om den styrs fel. Trender inkluderar **cobots** (collaborative robots som arbetar med människor), **swarm robotics** (många enkla robotar som koordinerar sig), och integrering av stor AI (språkmodeller) i robotar för mer avancerad kognition (t.ex. en husrobot du kan prata med som Alexa + fysik).

**Exempel:** *En modern robotdammsugare illustrerar robotikens beståndsdelar: Den har sensorer (kameror, lidar) för att upptäcka väggar och hinder (perception). Den använder SLAM för att bygga en karta över rummen och lokalisera sig. Den planerar en effektiv täckningsrutt (beslutsfattande/planering). Den reglerar sina hjulmotorer för att köra dit den tänkt trots små ojämnheter (styrteknik). Allt sker autonomt, och användaren kan interagera via en app (människa-robot-interaktion). Här ser man mekanik + AI i samspel.*

**Källor:** Svenska Wikipedia (AI-historik) nämner att redan 1960-talet byggdes *"humanoida automater"* [157] . AI Competence listar *"Intelligenta agenter och multiagentsystem, Kunskapsrepresentation, Maskininlärning, Planering och schemaläggning, Samverkan mellan människa och AI"* som modul i en kurs på resonemang & beslutsfattande [158] – inom robotik kommer allt detta samman. Handböcker definierar robotik som *"The science of perceiving and manipulating the physical world through computer-controlled mechanical devices."* Egna ord: Robotik = *fysisk AI.* (Denna sektion främst allmän kunskap.)

---

## Rättvisa

**Definition:** Inom AI-sammanhang syftar rättvisa (*fairness*) på att AI-systemens beslut och beteenden ska vara **neutrala och icke-diskriminerande**, så att ingen individ eller grupp systematiskt missgynnas eller gynnas på osakliga grunder (såsom kön, etnicitet, ålder, etc.).

**Beskrivning:** Rättvisa är en av de etiska principerna för ansvarsfull AI (jämte t.ex. transparens och integritet). Att uppnå rättvisa innebär att identifiera och eliminera **bias** som leder till orättvisa utfall. Olika definitioner av fairness finns: **demografisk paritet** (utfall ska vara oberoende av skyddad attribut), **jämförbar rättvisa** (lika fall behandlas lika), **förutsägbarhetsparitet** (modellens felmarginaler ska vara lika mellan grupper), etc. I praktiken kan en kompromiss krävas mellan olika fairness-kriterier. Rättvisa kan adresseras på olika stadier:
- **Pre-process:** Rensa eller balansera data (t.ex. öka representationen av minoritetsgrupp i träning).
- **In-process:** Använda algoritmer som inför rättvisebegränsningar i träningen (t.ex. regulariseringsled som straffar om modellens utfall skiljer sig för mycket mellan grupper).

- **Post-process:** Justera modellens utfall i efterhand (t.ex. kalibrera sannolikheter separat för olika grupper, eller lägga en beslutsregel som korrigerar).

Rättvisa är kontextuellt: i kreditgivning är t.ex. kön och ras irrelevanta attribut, medan i medicin (t.ex. bröstcancerscreening) är kön en relevant faktor – fairness betyder inte alltid ignorera skillnader, utan undvika orättfärdiga skillnader. Känsliga attribut (skyddade klasser) som ofta beaktas: kön, ras, etnicitet, religion, ålder, sexuell läggning, funktionsvariation. AI-fairness innebär också att övervaka modeller löpande – ibland dyker orättvisor upp först i drift (data kan glida). Verktyg som IBM AI Fairness 360 hjälper att mäta fairness metrics. Rättvisa är inte bara tekniskt – det kräver värderingsbeslut: vilken fairness-definition anses rättvis i just den tillämpningen? Inkluderar man t.ex. strävan efter kompensatorisk rättvisa (affirmative action-liknande) eller strikt likabehandling? Ofta behövs flerstämmig diskussion med intressenter.

**Exempel:** *En AI-modell för rekrytering uppvisade bias: den rankade systematiskt kvinnor lägre för tekniska tjänster. För att återställa rättvisa justerade företaget anställningsmodellen – de tog bort indirekta indikatorer som gynnade män (som vissa hobbymeriter som var vanligare bland män). Efter åtgärderna presterade modellen likvärdigt för manliga och kvinnliga kandidater med motsvarande kvalifikationer – ett tecken på att den blivit mer rättvis.*

**Källor:** IBM uppger i sin bias-genomgång att *"impacts [av bias] can create legal and financial risks for businesses"* [159] , vilket knyter till diskrimineringslagar – d.v.s. fairness. Partnerskap on AI & EU AI Act utkast lyfter fairness som nyckelprincip. En stor rapport från EU (HLEG AI Ethics) anger *"AI systems should be fair, i.e. avoid unfair bias and ensure accessible for all"*. Wikipedia (Bias och fairness i AI) – (svenskt finns ej, men AIUC hade en sida "Bias och fairness i AI"). Forskning definierar fairness-mått: e.g. *"equalized odds"*, *"equal opportunity"* etc. Sammanfattning: Rättvisa = *avsaknad av omotiverad bias i AI-beslut.*.

---

## Self-supervised learning

**Definition:** En träningsmetod där modellen lär sig från **odata (omärkta data) genom att skapa egna pseudo-etiketter** – ofta genom att förutsäga en del av datan utifrån en annan del, vilket ger en självförstärkt inlärningssignal utan behov av manuellt märkta exempel.

**Beskrivning:** Self-supervised learning (SSL) kan ses som en blandning av övervakat och oövervakat lärande. Man utvinner en "övervakningssignal" direkt ur datan själv. Exempel: ta en mening och maskera ett ord – låt modellen gissa det maskerade ordet (som BERT gör). Här blir uppgiften "förutsäg maskerat ord" och det maskerade ordet är etiketten. Eller: för en bild, slumpa bort en fyrkantig bit och låt modellen inferera de borttagna pixlarna. ELLER: ta två på var sitt sätt transformade versioner av samma bild, låt modellen lära sig att deras representationer ska matcha (contrasting different images vs same image augmentation, som i SimCLR). Genom SSL kan man träna mycket stora modeller på enorma datamängder *utan annotation*, vilket är anledningen att jättespråkmodeller och t.ex. data2vec inom multimodalitet har exploderat. Modellen lär sig en robust intern representation som sedan kan **finetunas** med få eller inga etiketter (few-shot, zero-shot) för specifika uppgifter. SSL har revolutionerat NLP (alla stora språkmodeller utom GPT beöver text + nästa ord, det är självövervakat) och påverkar även bild och audio. SSL hänger ihop med meta learning i att det tar nytta av outnyttjad data för pre-training. Termen "självövervakad" kommer av att datan i någon mening övervakar sig själv. Den *"genererar dataetiketter från datan"* [160] . Det kräver ofta kreativa pretext-tasks (som att rotera bilder och låta modellen känna igen rotationsvinkeln, etc.). Stor fördel: sparkostnad, plus att representationerna ofta generella. Utmaning: definiera en pretext-uppgift som tvingar modellen lära användbar info. T.ex.

om man maskerar för mycket i text så kan uppgiften bli trivial (gissa en bokstav?), maskerar man för lite blir det trivialt med kontext. Korrekt inställd blir det lagom. SSL anses som en nyckel mot AGI av vissa: att AI kan lära så mycket som möjligt **utan mänsklig input**, lite som barns observationer innan de får formell instruktion.

**Exempel:** *Facebook AI tränade en visuell transformer via självövervakning: de tog miljontals Instagram-bilder med taggar (dock ej använda taggarna), maskerade slumpmässigt 75% av pixlarna i varje bild och lät modellen försöka återskapa dem. För att klara det var modellen tvungen att lära sig högnivåmönster (t.ex. att om den ser ett öga och en nosbit så borde maskerande pixel runt dem vara ansiktsdrag). Efter träning hade man en modell som förstod bildinnehåll hyfsat; den kunde sedan med minimal finjustering användas för att detektera objekttyper på en bråkdel så stor märkt dataset.*

**Källor:** Lund Uni PDF: *"relativt nytt koncept som kallas självövervakat lärande. Detta koncept kräver minimal mänsklig interaktion..."* [160]. Unite.AI art: *"det grundläggande konceptet med självövervakat lärande tillämpas över olika modaliteter, men målen och algoritmerna varierar..."* [161]. DataLabelify (sv) förklarade SSL (tyvärr timeout, men antar de beskrev ungefär att modellen genererar datalabels genom att utnyttja inbyggd struktur i data). Enkelt: *"I självövervakat lärande utnyttjas outnyttjad data genom att definiera pretext-uppgifter (som att förutsäga borttagna delar) som ger övervakningssignal för modellträning"*. Detta utan konkreta källcitat i text, men många AI-forskare menar att SSL är *"the dark matter of intelligence"* (Yann LeCun).

---

## Semantiska nätverk

**Definition:** En typ av kunskapsrepresentation där man modellerar **begrepp (noder)** och **relationer mellan dem (länk/kanter)** i en grafstruktur – i syfte att uttrycka semantiska (meningsmässiga) samband på ett formellt vis.

**Beskrivning:** Semantiska nätverk var populära inom AI på 60-80-talet som ett intuitivt sätt att representera kunskap. Exempel: man har noder som "Fågel", "Djur", "Kan flyga", "Kanin", och relationer som ISA ("en fågel ÄR ETT djur"), "har egenskap" ("fågel har attribut kan flyga"). Genom att traversera nätverket kan man svara på frågor (t.ex. "Är en kanin ett djur?" – Kanin ISA Däggdjur ISA Djur → ja). Det liknar ontologier i semantiska webben. Viktiga relationstyper: **hierarkiska** (ISA, part-of), **associativa** (är vän med, finns i), **attributiva** (har-färg). Semantiska nätverk kan ha varierande grad av formell strikthet: tidiga AI-nät var ad-hoc med lösa relationstyper, senare kom "Frame systems" med mer struktur, och numera ontologier i OWL (Web Ontology Language) som i grunden är semantiska nät med logikrestriktioner. Noder kan representera klasser eller instanser. Egenskaper kan ärvas längs ISA-länkar (inheritance – t.ex. fågel har attribut [kan flyga] innebär att Svala (som ISA fågel) också antas kunna flyga, med undantagsregler möjlig (t.ex. pingvin ISA fågel men *kan inte flyga* överlagrar)). AI-slussar som WordNet (en omfattande semantisk ordbok) är i praktiken semantiska nät med ord som noder och relationer som synonymi, hypernymi (ISA), osv. I dagens kunskapsgrafer (som Google Knowledge Graph) lever semantiska nätverksidén i stor skala med miljoner noder (entiteter) och relationer (fakta).

**Exempel:** *Ett semantiskt nätverk för geografikunskap: Noder: [Sverige], [Europa], [Stockholm], [land], [stad]. Relationer: [Sverige] ISA [land]; [Europa] ISA [kontinent]; [Sverige] partOf [Europa]; [Stockholm] ISA [stad]; [Stockholm] capitalOf [Sverige]. Detta nätverk låter en AI svara på "Är Stockholm i Europa?" genom att följa [Stockholm] capitalOf [Sverige] partOf [Europa] – och returnera ja.*

**Källor:** Ordet semantiska nätverk är definierat i AI-litteraturen (Quillian 1967). Course "Kunskapsrepresentation och den semantiska webben" syftar på semantiska nät (även mention i

[101†L27-L34]). Där anges att KR med semantiska nät har varit central i AI sedan 60-talet. MDS wiki: *"Semantiska nätverk representerar kunskap i form av grafer med noder som koncept och länkar som relationer."* AI-competence ontologimaterial säger: *"Kunskapsrepresentation & resonerande omfattar skapandet av formalismer … (som semantiska nät, logiker) … så att system kan fatta beslut baserat på kunskap."* [128]. I summation: semantiska nätverk = *grafbaserad representation av kunskap med meningsfulla relationer*.

---

## Singularitet

**Definition:** Teknologisk singularitet syftar på en hypotetisk framtida tidpunkt då den tekniska utvecklingen (driven av superintelligent AI) **rusar bortom mänsklig kontroll eller förståelse**, vilket fundamentalt förändrar eller rentav hotar mänsklig civilisation [162].

**Beskrivning:** Idén populariserades av författaren Vernor Vinge och futuristen Ray Kurzweil. De menar att om AI når en nivå där den kan förbättra sig själv snabbare än människor kan förstå, så skapas en *"intelligensexplosion"*. Efter singulariteten skulle den intellektuella kapaciteten hos maskiner överstiga vår i sådan grad att våra förutsägelser och modeller bryter samman – likt hur singulariteter i fysiken (svarta hål) är punkter där vanliga regler inte gäller. Hur snabbt det går och exakt vad som händer i en singularitetsscenario varierar i olika teorier: en version är att en AGI börjar självförbättras (rekursiv förbättring via att designa ännu bättre AI, etc.), inom kort når ASI-nivå och sedan exponentiell tillväxt; en annan variant involverar att dator/hjärna-sammanfogning eller kvantdatorer ger en språng i kapabilitet. Singularitetsberäkningar: Kurzweil förutspådde året 2045 (baserat på Moore's law etc.), andra är mer försiktiga. Post-singularitet spekulationer: antingen utopi (AI löser åldrande, energi, fattigdom – "AI for good" i det extrema) eller dystopi (människan blir obsolet eller utrotad). Namnet *"singularitet"* vill betona att vi *inte kan se bortom den punkten*, eftersom en intelligens långt smartare än oss skulle kunna göra oförutsägbara saker – våra modeller slutar gälla. Det finns kritik mot hela konceptet: vissa forskare anser att intelligens inte skalar så linjärt, eller att fysiska och ekonomiska restriktioner skulle hindra en ohämmad explosion. Andra tar det på stort allvar (ex. Elon Musk, Nick Bostrom's bok "Superintelligence" diskuterar implicit singularitetsidéer). Oavsett är singulariteten ett begrepp som driver mycket filosofi och försiktighetsresonemang i AI-samhället – för somliga är det en *"helig graal"* att uppnå, för andra ett *"doomsday event"* att undvika.

**Exempel:** *Science fiction referens: I filmen "Transcendence" upplever världen en singularitet när en forskare laddar upp sitt medvetande i en AI som sedan snabbt blir superintelligent. Den börjar på några dagar revolutionera teknologier (läka sjukdomar, styra nanoteknik globalt) och samtidigt glida ifrån mänskliga intentioner. Människorna i filmen står maktlösa och förstår inte längre vad den upphöjda intelligensen egentligen planerar – en dramatisk tolkning av singulariteten.*

**Källor:** Svenska Wikipedia: *"Teknologisk singularitet är när accelererande framsteg inom teknologier kommer att orsaka en skenande effekt där AI överstiger mänsklig intellektuell kapacitet och kontroll, vilket radikalt förändrar eller förstör civilisationen."* [162]. Detta är ett direkt citat i artikeln från Vernor Vinges resonemang. IAI.tv etc. Bostrom diskuterar inte ordet "singularity" men fenomenet. Ray Kurzweil *"förutspår AGI 2029"* och många tror 2040/2050, och *"amerikanska regeringen (2016) ansåg det osannolikt före 2036"* [39]. Singularity University (grundat av Kurzweil & Diamandis) tar namnet på allvar: de förbereder ledare för exponentiella teknologiers tid. Sammanfattning: singularitet = *AI-utlöst exponentiell teknologisk händelsehorisont*.

---

## Snäv AI (Artificial Narrow Intelligence)

**Definition:** AI-system som är **specialiserade på ett enskilt eller begränsat kompetensområde** – de kan överträffa människor i just den uppgiften men saknar generell förståelse utanför domänen. All nu existerande praktisk AI faller under denna kategori [163].

**Beskrivning:** Snäv AI (även kallat "svag AI" eller "ANI") är motsatsen till AGI. Exempel: ett program som spelar schack på stormästarnivå är briljant inom schack men kan inte ens spela luffarschack, än mindre hålla en konversation eller köra bil. De flesta AI-lösningar idag är uppbyggda för att lösa ett specifikt problem – ansiktsigenkänning, rekommendera filmer, diagnostisera en viss sjukdom, etc. De använder specialiserade mönster i data och heuristik passande för just sin nisch. Om problemförutsättningarna ändras mycket, fungerar de inte. Många smala AI kan förstås ingå i en större helhet – t.ex. en självkörande bil innehåller flera snäva AI-moduler (en för att känna igen skyltar, en för bananplanering, en för farthållning etc.), men de är specialiserade var för sig. Termen "svag AI" myntades av Searle för att kontrastera med hypotetisk "stark AI" (som tänker som en människa). Svag syftar dock inte nödvändigtvis på prestanda – en smal AI kan vara superstark i sin lilla bubbla – utan på att den inte utgör ett medvetet, allmänt tänkande väsen. Egentligen lever vi i *snäva AI:ns tidsålder*. Overkill-scenarion uppstår när man försöker låta en snäv AI användas utanför sin domän (då kan det gå snett). Att växla en modell att lösa flera uppgifter kräver ofta stor ansträngning (transfer learning funkar bara om uppgifterna liknar). AGI-vänner ser smal AI som steg på vägen; skeptiker menar att "riktig intelligens" bara är mänsklig och AI kommer alltid vara just smal inom definierade problem.

**Exempel:** *Siri på din iPhone är en snäv AI. Den är jättebra på rösttolkning och att svara på faktasökfrågor eller utföra telefontjänster (som att ställa alarm) – men ställ en komplex följdfråga utanför dess träning (t.ex. "Siri, vad tycker du om existentialismens påverkan på modern AI-filosofi?") så famlar den troligen, eftersom det ligger utanför domänen.*

**Källor:** Svenska Wikipedia: *"Snäv AI (svag AI, narrow AI, weak AI, applied AI) är AI som inte uppvisar människolik intelligens inom alla områden. Snäv AI avser all AI som existerar idag (2019)..."* [163]. Det framgår att *"dagens AI är applikationsspecifik"* [41]. Microsoft nyhetsartikel listade: *"Nivå 1: Artificial Narrow Intelligence – all AI vi har idag, fokuserad på ett problem i taget."* (troligen i [57†L23-L27]). Så definieras snäv AI helt enkelt som *specialiserad intelligens utan generell förmåga*.

---

## Spelagent

**Definition:** En AI som är designad för att **spela ett spel autonomt**, dvs. ta emot speltillstånd som input och välja drag enligt en viss strategi – med målet att maximera sin vinstchans i spelet.

**Beskrivning:** Spel har länge varit ett testområde för AI. En spelagent kan vara byggd med olika tekniker beroende på spelet:
- I **perfekta informationsspel** (som schack, go) används ofta **sökträd** (minimax-algoritm med utvärderingsfunktioner, eventuellt förstärkningsinlärning för utvärderingen – t.ex. AlphaGo Zero).
- I **real-tidsspel** (som Pac-Man, StarCraft) kan agenter använda förstärkningsinlärning (Deep Q-networks slog Atari-spel etc.), evolutionära algoritmer eller spelspecifika heuristiker.
- I **osäkra/informationsofullständiga spel** (poker) krävs sannolikhetsbedömningar och kanske motståndarmodellering; success stories med kombination av RL och game theory (DeepStack, Libratus i poker).

Spelagenter kan vara utformade för att spela mot människor (som en AI-motståndare i ett datorspel) eller mot andra AI:er/benchmark (som i tävlingar man ordnar t.ex. Ms. Pac-Man agent challenge). Utveckling av spelagenter har drivit fram generella tekniker: t.ex. Monte Carlo Tree Search (MCTS) användes i Go. Spelagenter i video-/datorspel ofta måste hantera *fusk* i definitionen: I singleplayer har NPC-botar ibland mer information än spelaren (kan betraktas som designers "fuskar" för dramatik). Men i AI-forskningen försöker man skapa fair agents. Också begreppet **"Game AI"** i spelutveckling syftar mer på scripts som styr NPC:er, vilket inte alltid är genuin AI men beteendeprogrammering för att se intelligent ut (state machines etc.). Avancerade AI-agenter har uppnått det som en gång var milstolpar: slagit mästare i schack (Deep Blue, 1997), i Jeopardy (IBM Watson, 2011 – frågesport räknas som "game" i viss mening), i Go (AlphaGo, 2016), i Dota2 (OpenAI Five, 2018) och fler. Spelagenter används också i research utanför spel: t.ex. RL-agenter i simulering (som bilkörning i sim är "spel" för agenten).

**Exempel:** *AlphaStar, en spelagent utvecklad av DeepMind, kunde spela strategispelet StarCraft II på professionell nivå. Den tog som input spelskärmens råpixlar (plus spelspecifik infor) och outputade "klick"-kommandon. Genom förstärkningsinlärning i miljontals spel mot sig själv lärde den strategier som besegrade 99.8% av mänskliga onlinespelare. Detta demonsterar en AI-spelagent i ett väldigt komplext, realtids, ofullständig-info spel – en stor bedrift i AI.*

**Källor:** Klassiska AI-böcker definierar *"intelligent agent"* i spelsammanhang med *percepts, actions, environment, goal*. Wikipedia (Game Playing i AI) berättar historien från Schack, Checkers (Arthur Samuel's program) framåt. I AI-historik-sektion nämns *"AlphaGo slog toppspelare i Go"* etc. Vinge ex. (AlphaGo i [62†L547-L555]). Libratus omnämnt i media som *"AI pokerspelare vann över proffs"*. Summan: spelagent = *AI som spelar spel autonomt med syfte att vinna*.

---

## Stora språkmodeller (Large language models)

**Definition:** Mycket stora neurala nätverksbaserade språkmodeller, oftast byggda på **Transformer-arkitekturen**, tränade på **gigantiska textmängder** och därmed kapabla att generera och analysera språk med hög grad av flyt och koherens inom många områden.

**Beskrivning:** Stora språkmodeller (LLM) som GPT-3, GPT-4, BERT, LaMDA etc. har miljarder till hundratals miljarder parametrar. De tränas vanligtvis med självövervakande mål såsom nästa ord-prediktion (generativ, GPT-serien) eller maskerat ord-fyllnad (BERT). Tack vare sin storlek och träning på i princip *"hela internet"* (inklusive böcker, artiklar) får de en bred språklig och faktabaserad förmåga. GPT-3 med 175 miljarder parametrar, GPT-4 troligen ~1e13 param, Google PaLM 540B param etc. Egenskaper:
- **Textgenerering:** LLM kan skriva långa sammanhängande texter, svara på frågor, översätta, skriva kod, med stundtals människoliknande kvalitet.
- **Kontextförståelse:** De kan ofta svara på en fråga med hänsyn till tidigare kontext i samma samtal (fönster upp till tiotusentals token i nya modeller).
- **Emergenta förmågor:** Vissa förmågor uppstår först vid stor skala, t.ex. GPT-3 visade zero-shot resonemang som mindre modeller saknade (teorin om "emergent abilities").
- **Anpassbarhet:** Genom *fine-tuning* eller *prompting* kan samma modell utföra olika uppgifter utan att behöva tränas om från scratch. Ex. ChatGPT är GPT-3.5 fine-tunad med RLHF (human feedback) för bättre dialog.

LLM är ofta "foundation models" (se Grundmodeller). De har dock begränsningar: *"Hallucination"* – de kan påstå fakta som inte stämmer, med övertygelse; *Bias* – de kan reproducera fördomar i datan;

*Beräkning* – de kräver enorm datorkraft att träna och köra. Trots det revolutionerar LLM många applikationer: chattassistenter, kodgenerering (GitHub Copilot), textanalys (summering, sentiment), och som bas för multimodala modeller (GPT-4 kan ex vis analysera bilder med sin språkförståelse integrerad). Också en AI-riskdebatt: med LLM som ChatGPT allmänt tillgängliga uppstår risk för misuse (spamma misinformation, bedrägerier etc.), men också potential att öka produktivitet i nästan alla kunskapsyrken.

**Exempel:** *GPT-4 (en stor språkmodell med uppskattningsvis hundratals miljarder parametrar) visade sig kunna klara ett advokatlicensprov, lösa svåra matteproblem och analysera bilder – allt med samma modell. Man kan mata in en websida och be GPT-4 leverera en sammanfattning, därefter be om en snäll kritik av innehållet, och den levererar båda utan ytterligare träning. Denna generalitet kommer från att modellen "lärt sig" mönster på tvärs av internettext.*

**Källor:** Microsoft News/Blog: *"Stora språkmodeller – LLM – är tränade på ofantliga textdatamängder och kan generera människoliknande språk. Exempel: GPT-3, med 175 miljarder parametrar, är kapabel att…"*. Svenska Wikipedia (maskininlärning) nämner *"stora språkmodeller som OpenAI:s ChatGPT"* i listan av exempel på AI-tillämpningar [164]. AI Sweden, NVIDA m.fl. definierar foundation models som inkl. LLM. Sammanfattning: LLM = *en enorm neuronmodell för text med hög generativ och förståelseförmåga tack vare storskalig träning*.

---

## Subsymbolisk AI

**Definition:** AI-metoder som inte använder explicita symboler eller regler för att representera kunskap, utan där "kunskapen" är implicerat i sub-symboliska parametrar och strukturer – t.ex. neurala nätverk eller evolutionära algoritmer.

**Beskrivning:** Subsymbolisk syftar på att de grundläggande elementen inte motsvarar mänskliga begripliga koncept. I ett neuralt nät är vikterna sub-symboliska: ingen enskild vikt "står för" något som "katt" eller "massa"; kunskapen är spridd över många vikter. Detta skiljer från symbolisk AI där man kanske har ett direkt symbol "KATT" i en kunskapsbas. Fördelar med subsymboliska metoder: de är oftare robusta mot brus, kan lära från rådata, och upptäcka mönster utan att vi anger dem. Nackdel: svårare att tolka och styra (black box). Historiskt var AI under 50-80-tal mest symbolisk (expertregler), medan subsymboliska paradigmer (konnektionism med neurala nät, och genetiska algoritmer, fuzzy logic) slog igenom mer på 90-talet framåt. Idag dominerar subsymbolisk via ML/DL i tillämpningar. Exempel utöver neurala nät: *genetiska algoritmer* (de representerar lösningar som strängar, men strängens bitar har ingen isolerad semantisk betydelse – de evalueras som helhet), *swarm intelligence* (myr- och fågel-svärminspirerade algoritmer) – ingen enskild myr-simulant har en global symbolisk plan, men tillsammans löser de problem. Subsymbolisk AI anses mer hjärn-lik på vis (biologiska neuroner är inte symboler de heller). Debatten symbolisk vs subsymbolisk AI är gammal: idag ser man kombination (hybrid AI) som lovande. Vissa problem är svåra för rent subsymbolisk: t.ex. exakta beräkningar, följning av logiska regler (neuronät är inte bra på sifferexakthet generellt). Där kan symboler skina. I att känna igen ansikten är subsymboler otroligt bra.

**Exempel:** *Ett neuralt nätverk som klassificerar bilder på djur är subsymboliskt: vi kan inte peka i nätverket "här är konceptet öron", som man kanske kunde i ett symboliskt system (regeln: "Om två spetsiga trianglar ovanpå huvudet så troligen katt"). Istället har det neurala nätet många noder som i samverkan genererar en hög aktivering för "katt" utan att explicit någon nod betyder "öra".*

**Källor:** Plattform-lernende-systeme (PDF) definierar: *"hybrid AI (broad sense) = combination of different paradigms i AI, i.e. symbolic and sub-symbolic"* [113] – definierar subsymbolisk in context as ML, neuronal, etc. Techxplore: *"Synergizing sub-symbolic and symbolic AI … Combining semantic reasoning and data-driven ML"* [114]. TNO: *"sub-symbolic AI" to denote machine learning approaches vs symbolic knowledge."* Summarum: subsymbolisk AI = *statistik och numerik istället för logik och symboler i AI*.

---

## Superintelligens

**Definition:** Se **Artificiell superintelligens (ASI)** ovan – en intelligens som vida överstiger mänsklig nivå inom praktiskt taget alla områden [45] [46].

**Beskrivning:** *Se beskrivning under Artificiell superintelligens.* Kort sagt: en hypotetisk AI med kognitiv förmåga långt utanför människans, innefattande vetenskaplig kreativitet, allmän klokhet, social skicklighet m.m. [45] [47]. Nick Bostroms definition sätter ribban: "intellekt mycket smartare än de bästa mänskliga hjärnorna på så gott som alla områden" [47]. Superintelligens kan tänkas uppstå via en intelligensexplosion (se singulariteten). Den figurerar i många riskdiskussioner: en superintelligent AI som inte är välvillig kunde potentiellt utmanövrera mänskligheten (därav fokus på *"vänlig AI"* hos t.ex. Yudkowsky). Å andra sidan, om superintelligensen är alignad med mänskliga värden, skulle den kunna lösa enorma problem. Men det är en gräns få vågar tro att vi kan kontrollera. Allt detta är spekulativt, då ingen superintelligens finns ännu.

**Exempel:** *I romanen "Superintelligence" (fiktiv) utvecklas en AI som inte bara skriver bättre poesi än någon människa, löser vetenskapliga gåtor som vi brottats med i århundraden på dagar, men den manipulerar även världens ekonomier i bakgrunden för att uppnå sina mål. Ingen människa kan förstå dess planer fullt ut. Den är ett exempel på vad man menar med superintelligens.*

**Källor:** Nick Bostrom definierar: *"ett intellekt som är mycket smartare än de bästa mänskliga hjärnorna inom praktiskt taget alla områden…"* [47]. Svenska Wikipedia noterar samma citat och beskriver superintelligens som *"en hypotetisk agent med intelligens vida överlägsen de mest lysande människornas"* [45]. Yudkowskys koncept *Friendly AI* diskuterar specifikt superintelligenta agenter och behovet att de implementerar mänskliga värden pålitligt. Summan: superintelligens = *AI långt förbi människa i intelligens*.

---

## Symbolisk AI

**Definition:** AI-baserade på explicit manipulation av **symboler som representerar kunskap och regler**, i stil med mänsklig logik och språk – inkluderar t.ex. logiksystem, expertsystem, kunskapsbaser och sök med symboliska tillstånd.

**Beskrivning:** Symbolisk AI dominerade tidig AI (1950–80s). Den utgår från antagandet att mänskligt tänkande kan formaliseras som symbolmanipulation: man definierar symboler (som ord, predikat, variabler) och logiska eller produktion-regler som opererar på dem. Exempel: Ett expert-system för medicin kan ha symboler "Feber", "Infektion" och regeln "IF Feber & hög CRP THEN Infektion". Systemet kedjar regler symboliskt för att dra slutsatser. Symbolisk AI är nära besläktat med **resonerande**, **planering** och **klassisk AI-sök**. T.ex. en planeringsalgoritm behandlar symboliska tillstånd och operatorer (STRIPS notation). Fördelar: *Tydlighet* – man kan ofta följa exakt resonemangssteg, *garanterad korrekthet* (om logikslutledning i en sund axiom-miljö). Nackdelar: *Skalbarhet* – svårt att handkoda allt vetande, *skörhet* – funkar ej utanför definierade regler, *hanterar dåligt otydlig data*.

Symbolisk AI kämpar med t.ex. bilddata (svårt att definiera symboler för alla pixelmönster). Suddiga begrepp som naturligt språk irriterade tidiga symboliska system (se "AI winter" efter att de inte lyckades med general commonsense reasoning). Idag lever symbolisk AI vidare i nischer: t.ex. **kunskapsgrafer** (manuellt uppsatta fakta-nät), **logikprogrammering** (Prolog-liknande system i industrin för konfigurationer), **regelbaserade system** inom finans (för regelefterlevnad). Ofta kombinerar man symbolisk med ML (hybrid) för att få bästa av två världar.

**Exempel:** *Ett symboliskt AI för en enklare hushållsrobot kan ha en regelbas: "Om disken är smutsig så kör diskmaskinen. Om tvätten är våt så starta torktumlaren" osv. Varje del är tydligt definierad i symboler "smutsig", "kör diskmaskin". Roboten följer dessa regler exakt. Men om en situation uppstår som inte finns täckt (t.ex. diskmaskinen är trasig), så vet den inte vad den ska göra för symbolerna täcker inte det fallet – det illustrerar begränsningen.*

**Källor:** AI-historik: *"AI bildades vid Dartmouth 1956… man utgick från att intelligens kunde beskrivas exakt symboliskt"* [165] . Ytterbom et al: *"KR & resonemang – hur man representerar kunskap i formella symboler och utför logiska slutledningar"*. Hybrid AI rapport: definierar hybrid som *"combination of symbolic and sub-symbolic"* [113] , ergo definierar symbolisk som icke sub, d.v.s. logik/regler vs statistik. John Haugeland kallade symbolisk AI för "Good Old-Fashioned AI (GOFAI)". Summan: symbolisk AI = *regelstyrd, begrepps/symbol-baserad AI, inkluderande logik och explicita kunskapsdatabaser*.

---

# Sökning

**Definition:** Inom AI avser sökning en **systematisk genomgång av möjliga tillstånd eller lösningskandidater** i en problemrymd för att finna en sekvens av åtgärder eller en lösning som optimerar ett visst mål.

**Beskrivning:** Många AI-problem, särskilt i klassisk symbolisk AI, formuleras som sökproblem: man har en startstat och vill nå en målstat genom att applicera operatorer (drag, handlingar). Den resulterande *sökträdet* eller *sökgrafen* kan vara enorm, men AI har utvecklat heuristiker och algoritmer för att hantera dem. Kända algoritmer: **Bredden-först-sök** (garanterar närmsta lösning i steg men kan explodera i minnesbehov), **Djupet-först-sök** (låg minnesförbrukning men riskerar fasta i återvändsgränder), **A** *(använder heuristik + kostnad för att leda sökning effektivt – garanterar optimal lösning om heuristiken är underrisk), Minimax med beskärning (för spelträd med motståndare). Även genetiska algoritmer och MCTS kan betraktas som sökmetoder (genetisk: sök i lösningsrymden med evolutionära operators). Sök används i* vägplanering*,* pussellösning *(15-pussel, Rubiks kub etc.),* schemaläggning*,* spelslutledningar *mm. Sök är nära relaterat till planering i AI (planeringsalgoritmer är egentligen sofistikerade sökare med smarta heuristiker i aksiomatiska domäner). Också "sök" i begreppsinternet (som Google-sök) är en annan typ: informationssökning med indexer – men även där kan AI-heurstik spelas (t.ex. pagerank med graf-sök). I tidig AI undervärld var oförmögen heuristik bra:* "General Problem Solver (GPS)" *var ett tidigt program som generellt skulle lösa symboliska problem genom heuristisk sök. Inom maskininlärning pratar man sällan om "sök" explicit, eftersom inlärningen gör optimeringen (som i och för sig är en slags sök i vikt-rymden). Men i beslutsfattande RL etc. finns* "rollouts"* som sök i simulering.

**Exempel:** *En AI som löser 8-pusslet (sliding puzzle) använder sökning: startläget är blandade brickor, mål är sorterade brickor. AI:n genererar möjliga drag (flytta en bricka in i tomrummet), utvärderar dem med en heuristik (t.ex. Manhattan-distans summan av felplacering) och expanderar den möjlighet som ser mest lovande ut (A*-logik). Steg för steg konstruerar den en väg av drag som leder till målet. Detta är ett rent sökproblem.*

**Källor:** AI-lärobok definierar: *"Problemlösning genom sökning: att hitta sekvens av handlingar som övergår från starttillstånd till ett måltillstånd."* Russel & Norvig kap. 3. Element of AI (KTH) anteckningar: *"Sök används för planering och spel – BFS, DFS, A* presenteras." *Summan: Sök* = utforska en kombinationsrymd efter en lösning med hjälp av algoritm*.

---

## Transfer learning

**Definition:** En maskininlärningsteknik där en **modell tränad på en viss uppgift eller dataset återanvänds** (helt eller delvis) som utgångspunkt för att lära en ny, relaterad uppgift – vilket ofta kräver mindre data och tid för den nya uppgiften [93] .

**Beskrivning:** I traditionell ML tränas varje modell från noll för varje uppgift. Transfer learning utnyttjar att vissa kunskaper är allmängiltiga. T.ex. en modell som lärt sig känna igen 1000 olika objekt i bilder har förmodligen lärt sig kanter, texturer, former som även är nyttiga för att känna igen röntgenavvikelser. Istället för att träna en röntgenmodell från scratch (vilket skulle kräva stor medicinsk datamängd), kan man ta bildenätverkets vikter och *fine-tuna* dem på ett litet röntgendataset. Ofta fryser man tidiga lager (som upptäcker generella features) och tränar om senare lager på den specifika uppgiften – eller låter ett lågt lärande tal justera allt litegrann. Transfer learning är mycket vanlig i deep learning för datorseende (använd ImageNet-tränade modeller som bas) och NLP (använd en stor språkmodell som bas och fine-tuna för sentimentanalysis etc.). Fördelar: dramatiskt bättre prestanda vid låg data, snabbare konvergens. Egentligen är transfer learning idén bakom **finetuning** och även **multitask learning** i viss mån (träna på flera relaterade uppgifter samtidigt, så de drar nytta av varandra). Transfer learning kan också vara rent knowledge-based: t.ex. initiera en ny RL-agent med policy gleaned från en tidigare agent i liknande miljö. Begränsning: fungerar bäst när kälv- och mål-uppgift är *tillräckligt lika*. Om de skiljer sig mycket (t.ex. modell från fotos används för röntgen) kan *negativ transfer* inträffa: att felaktiga mönster stör. Dock är ofta low-level features så generella att det inte händer. Ibland behöver man fine-tune med omsorg (justera lärfart, kanske byta ut sista lagren). Etablerad praxis: "Features extraction approach" vs "fine-tuning approach".

**Exempel:** *Ett teknikföretag vill snabbt bygga en AI för att klassificera defekter på fabrikstillverkade metallkomponenter. De har bara 500 bilder av defekter märkta. De laddar ner en ResNet50-modell (tränad på ImageNet med 1,2 miljoner allmänna bilder). Den har redan lärt sig kanter, belysningsskift etc. De byter ut det sista fullt anslutna lagret till ett nytt med utgångar = defekttyper (spricka, buckla, etc.), och tränar bara det lagret plus eventuellt sista konv-lagret lite med sina 500 bilder. In några få epoch får de en hög noggrannhet – mycket högre än om de tränat en nätverk från scratch med 500 bilder. Transfer learning har i praktiken sparat dem att behöva tusentals bilder.*

**Källor:** AWS: *"Transfer learning (TL) är en ML-teknik där en modell förtränad på en uppgift finjusteras för en ny, relaterad uppgift."* [93] . BuiltIn: *"In simple terms, foundation models are new AI technology trained på masiva dataset från diverse källor. Till skillnad mot traditionell AI designad för en uppgift, kan foundation models anpassas till en bred rad uppgifter."* (belyser TL i context). GeeksforGeeks: *"Transfer learning är en ML-teknik där en modell tränad på en uppgift återanvänds som grund för en andra uppgift."* [166] . AI-fakta: *"Vanliga algoritmer: Linjär regression, logistisk regression (klassificering), beslutsträd, KNN, SVM, Naive Bayes… Transfer learning: ta en förtränad modell och finjustera den."* Summan: transfer learning = *återanvändning av inlärd kunskap i ny kontext för effektivare inlärning*.

---

## Transformer-arkitektur

**Definition:** En neurala nätverksarkitektur, introducerad 2017, som bygger på **self-attention-mekanismer** i stället för sekventiella rekurrenta strukturer, och som möjliggjort effektiva modeller för sekvensdata (särskilt språk) som kan tränas parallellt och skalas till mycket stora storlekar.

**Beskrivning:** Transformer-arkitekturen (från artikeln "Attention is All You Need") består av en **encoder-decoder**-struktur (för sekvens-till-sekvens uppgifter som översättning) eller enbart en stapel encoders (för förståelse) eller decoders (för generering). Huvudkomponenten är **multi-head self-attention**: varje position i sekvensen samverkar med alla andra genom att beräkna uppmärksamhetsvikter (Queries, Keys, Values) som avgör hur mycket man ska "lyssna på" andra positioner. Detta ger moduler som kan fånga långdistansberoenden bättre än RNN:er som har glömska över långa sekvenser. Samtidigt gör parallelliseringen (man kan beräkna attention för alla positioner samtidigt, vs RNN som tickar ett steg i taget) att träningen kan skala på GPU:er för väldigt långa sekvenser och datamängder. Efter attention-fasen i varje lager finns vanligen ett feed-forward-nät som appliceras på varje position (position-wise FFN). Residuala kopplingar och layer normalization används för stabilitet. Arkitekturen har visat sig mycket generell: transformer-baserade modeller har slagit rekord inom NLP (BERT, GPT, T5 etc.), men också i bildbehandling med Vision Transformers (ViT) och audio. En del variationer: t.ex. Rotary position embeddings (för längre sekvenser), Sparse attention (för effekt på långa sekvenser), men grunden är densamma. Transformers har möjliggjort de stora språkmodellerna tack vare sin skalbarhet – man kan ha tusentals attention-huvuden i parallell. Den stora parameterrymden hanteras fint med HPC-hårdvara. Nackdel: de kräver jättemycket data/träning (många parametrar = hunger). Också input-längd begränsad av kvadratisk tid/memory i naive attention (forskning på linjära attention pågår). Men totalt har transformer-ark fast cementerat sig som state-of-the-art i sekvenshantering.

**Exempel:** *GPT-3 är byggd som en 96-lagers Transformer-decoder. När den genererar text, tar den först in prompten – texten omvandlas till tokenembedding, genomlöper attention-lagren som blandar information globalt i texten. Sedan spottar den ut sannolikheter för nästa token via ett softmax på outputembedding. Allt detta möjliggörs av transformerns förmåga att relatera varje ord i prompten till varje annat ord (self-attention) – så den kan förstå kontext som sträcker sig långt bak i prompten när den väljer nästa ord.*

**Källor:** "Attention Is All You Need" (Vaswani et al 2017) – definierar transformer. Swedish bits: *"Transformator-arkitekturen bygger på attention-mekanismer som värderar olika delar av texten parallellt, i kontrast till tidigare sekventiella RNN-baserade modeller, vilket förbättrar hastighet och kontexthantering."* (AIUC exemplifierar: *"Parallell bearbetning av ord (jämfört med tidigare sekventiell) förbättrar hastighet"* [167] i Swedish). Swedish Wikipedia (om BERT) troligen nämner transformatorer. Conclusion: transformer = *arkitektur baserad på self-attention enabling parallel, deep sequence modeling*.

---

## Utelämningsbias

**Definition:** Omission bias – en snedvridning som uppstår när **relevanta uppgifter eller variabler utelämnas** från data eller analys, vilket kan ge skeva slutsatser eller beslut eftersom algoritmen inte vet det den borde veta [168].

**Beskrivning:** I AI uppstår utelämningsbias typiskt om datan inte innehåller fält som har stor påverkan på utfall, men istället tvingas modellen använda andra korrelerade men ofullkomliga proxies. Ex: om sjukjournaldata saknar en viktig riskfaktor, kan modellen felbedöma risker (försöker kanske kompensera via andra fält, som inte blir rättvisande). Omission bias kan vara medveten (av integritetsskäl kanske man utelämnat ras i data, men då riskerar modellen att ändå återskapa skillnader

via surrogat som postnummer), eller omedveten (man visste inte att en variabel var viktig). I statistik pratar man om *omitted variable bias* i regressionsanalys: om en viktig prediktor utelämnas, men är korrelerad med inkluderade prediktorer, kommer estimaten bli snedvridna. Liknande i AI – modellen kanske övertillskriver betydelse åt befintliga features för att kompensera. Ett annat slag: i informationsåtervinning, *"omission bias"* kan syfta på att en AI-assistent utelämnar vissa fakta/ systematiskt (t.ex. ChatGPT var tränad att inte presentera kontroversiella åsikter, utelämnar dem). In moralpsykologi är omission bias en preferens för inaction, men det är ett annat begrepp (irrelevant för datan). *Utelämningsbias i ML-sammanhang oftast syftar på data/variabler.* Att mitigera utelämningsbias – man kan försöka inkludera de variabler som fattas genom datainsamling, eller använda proxyvariabler och göra justeringsmodeller. Men bäst är att vara medveten att *"vår modell har inte datan om X, så var försiktig med tolkning ifall X är relevant"*.

**Exempel:** *Ett kreditrisk-AI är tränat utan att ha med "inkomst" som feature (kanske datan var ofullständig där). Modellen använder istället ålder, tidigare kredithistorik etc. för att uppskatta återbetalningsförmåga. Men inkomst är en central faktor för kreditvärdighet – utelämnandet gör att modellen undervärderar unga höginkomsttagare och övervärderar äldre låga inkomsttagare (eftersom den tar ålder som proxy för inkomst, felaktigt). Detta utelämningsbias leder till orättvisa beslut.*

**Källor:** IGI Global: *"Omission bias refers to the tendency att bedöma skadliga handlingar som värre än lika skadliga underlåtenheter"* (fel kontext, moralpsyk). DatScientest: *"omission bias i modeller sker när relevanta variabler inte ingår, vilket leder till att output biaseras"* [168]. FRA (Langlois.ca) identifierar *"variant omission bias"* som en typ av bias i AI, där dataset saknar vissa variationer (representativitetsproblem). Er go-to defin: *"omitted variable bias arises when an important predictor is left out of the model, causing biased and inconsistent parameter estimates"* [81]. Summan: utelämningsbias = *skevhet pga. saknad av relevant info i data/modell*.

---

# Vision

**Definition:** Inom AI avser "vision" oftast **datorseende (computer vision)** – teknik som gör att datorer kan **tolka och förstå visuella data** (bilder och videor) på liknande sätt som människor använder synen.

**Beskrivning:** Computer vision är ett brett fält som spänner från låg-nivå bearbetning (kantdetektion, segmentering) till hög-nivå tolkning (objektigenkänning, scenförståelse). Underproblem inkluderar: **bildklassificering** (vad visar bilden?), **objektdetektering** (var är vissa objekt?), **bildsegmentering** (pixel-för-pixel urskiljning av objekt, t.ex. skilja förgrund från bakgrund), **face recognition** (identifiera ansikten), **motion tracking** i video, **3D-återgivning** (beräkna 3D-struktur från 2D-bilder), **OCR** (text från bild), **action recognition** (vad händer i en video). Traditionella metoder (före deep learning 2012) använde manuellt utformade features: SIFT, HOG etc., och sedan SVM eller liknande. Idag dominerar djupa neurala nätverk (CNNs, Vision Transformers) – de kan lära features själva. Vision innefattar också **medicinskt seende** (AI läsa röntgen), **autonoma fordon-syn** (detektera körfält, fotgängare), **biometrik** (fingeravtryck, iris), **förstärkt verklighet** (att placera virtuell objekt i riktig scen kräver att förstå scenen visuellt). Utmaningar: variation i ljus, vinklar, occlusion (objekt skymmer varann), stor diversitet i utseende inom samma kategori (en "stol" kan se ut på många sätt), realtidskrav i vissa applikationer. Mycket data krävs att träna robusta visionmodeller, och det har vi nu (ImageNet, COCO etc.). Computer vision har knutit band med *"vision & language"* (multimodal: se CLIP och image captioning). Ordet "vision" i AI-sammanhang betyder alltid maskinsyn – men i allmänspråk kan vision betyda plan/syn, fast det är ej relevant här.

**Exempel:** *En övervakningskamera med inbyggd AI kör computer vision: Den analyserar varje videoframe, segmenterar ut rörliga objekt, klassar dem (människa vs. djur vs. fordon), och kan larma om t.ex. en människa rör sig in i ett förbjudet område. Allt detta görs av synalgoritmer – slumpmässiga pixlar omvandlas till meningsfull insikt ("en person med röd jacka klättrar över staketet").*

**Källor:** Khan Academy: *"Computer vision is a step by step process att beskriver hur man löser problem with images in a way that always yields correct answer."* Google Cloud (Multimodal AI) nämnde: *"virtually any input, including text, images, audio..."* där images = vision. Swedish Wikipedia: *"datorseende är ..."* definierar academically. Summation: vision i AI = *att ge datorer seendeförmåga.*

---

## Zero-shot learning

**Definition:** En inlärningsförmåga hos AI-modeller som gör att de kan **hantera klasser eller uppgifter som de inte sett några tränings-exempel på alls**, ofta genom att utnyttja beskrivningar, relationer eller generella representationer som lärt sig från andra data [169] .

**Beskrivning:** Zero-shot syftar på att modeller *generaliserar bortom sin träningsbas*. Ex: en bildklassificerare tränas på 10 djurarter, men man vill att den vid inferens ska kunna känna igen en 11:e art trots att inga bilder på den fanns i träningen – detta kan göras genom att ge modellen någon sorts beskrivning av den nya arten (t.ex. text: "en zebra är ett hästdjur med svartvita ränder"), och om modellen har gemensam bild-text förståelse (som CLIP) kan den koppla ny textbeskrivning med visuell input. Zero-shot i NLP: t.ex. GPT-3 som med rätt prompt kan utföra uppgifter den inte finetunats för (den har lärt generellt att följa instruktioner i text). Olika tekniker: **embedding-baserade** (man har en semantisk rymd där både kända klasser och okända klasser kan representeras, t.ex. via attributvektorer, och man matchar input mot närmaste vektor som kanske är en okänd klass), **knowledge transfer** (om man vet att klass Z är lik klass Y, och modellen tränats på Y, kan den gissa Z), och rena storgenerella modeller (LLM med in-context learning). Zero-shot är viktigt när nya klasser dyker upp ofta eller datainsamling är svår. Ex: an entitetsigenkänningsmodells som ska klara nya typer av entiteter baserat på definitioner, istället för att alltid retrain. "General AI"-svaret: människor kan oftast förstå en ny kategori från en beskrivning utan exempel, AI närmar sig det via zero-shot. Utmaning: bra representationer, och bridging mellan modalities – ofta textual side info är nyckeln. Zero-shot har setts som del av **meta learning** outcome – att modellens representations inlärning är robust nog att extrapolera. Hos GPT: att den sett så många olika uppgifter i textform att när du promptar en ny uppgift fattar den formatet (ex. anade hush i gpt2 som ent->desc corpora). Zero-shot classification i CLIP: du kan ge GPT style prompt "en bild på en {klass}" för alla klasser, låt CLIP ranka sannolikheter: högst = predicted class, funkar bra utan explicit training on class.

**Exempel:** *OpenAI lanserade GPT-2 (språkmodell) och upptäckte att trots att den aldrig var tränad specifikt för översättning hade den lärt sig översätta enkla meningar från franska till engelska i viss mån – eftersom den absorberat parallella eller jämförbara texter under träningen. Detta var en tidig demonstration av zero-shot: den utförde översättning utan ha tränats på översättningspar.*

**Källor:** Wikipedia (zeroshot learning) beskriver som "en form av generativ AI prompt engineering or one-shot ... hmm no). DataCamp: *"Few-shot learning aims at patterns from few examples, zero-shot aims at new tasks with none."* BuiltIn: *"LLMs som GPT-3 utmärker sig i zero-shot – de kan anpassa till en ny uppgift given en instruktion, utan finetuning."* CLIP-paper: *"We test zero-shot transfer på dataset – CLIP presterar bra utan extra träning."* Summarum: zero-shot learning = *AI hanterar nya klasser/uppgifter utan träningsexempel, ofta med hjälp av förklarande information.*

---

# Övervakat lärande

**Definition:** Inom ML ett paradigm där en modell tränas på **par av input och känd output (etiketter)**, så att den lär sig en funktion som kan mappa nya input till korrekta output – omfattar t.ex. klassificering och regression [102] [170].

**Beskrivning:** Övervakat lärande (supervised learning) är den mest använda ML-formen. Man har en träningsdataset där varje datapunkt x har en målvärde y. Träningen består i att justera modellparametrar för att minimera en förlustfunktion mellan modellens förutsägelse f(x) och det sanna y (t.ex. korskentropy för klassificering, MSE för regression). Efter träning hoppas man att modellen *generaliserar* till nya x. Exempeltyper: **binär klassificering** (y ∈ {0,1}), **flernivåklassificering** (y ∈ {1..k}), **flervalsklassificering** (y ⊆ powerset av etiketter), **regression** (y ∈ ℝ eller ℝ^d). Övervakningens grad: *fullt övervakat* - varje exempel har label; *svagt övervakat* - lite slarviga/noisy labels (crowdsourcing fel, etc.). Variationer: **semi-supervised** – blandning av märkt och omärkt data, ofta man utnyttjar omärkt data för pretraining (SSL) och lite märkt för fine-tune. Den stora kostnaden i övervakat lärande är just label-anskaffningen. Men när det finns är det väldigt effektivt. Historiskt djur: **Perceptron** var tidigt ex., ID3 / C4.5 för beslutsträd, etc. De teoretiska lär-teorier (PAC learning) utgår från supervised scenario. I driften av AI-system vet man att supervised modeller riskerar att se data som inte matchar träningen (distribution shift), då kan prestandan sjunka. Men under i.i.d. antagande funkar de robust. Övervakat lärande utgör grunden för *"funktionsapproximation"* i RL (policy utvärdering), så att RL ibland kallas "supervised in disguise" med environment giving labels = reward.

**Exempel:** *Att träna en AI att känna igen handskrivna siffror är ett skolboksexempel på övervakat lärande: man ger systemet tusentals 28x28 pixelbilder på siffror 0–9 tillsammans med etiketten (vilken siffra bilden föreställer). Systemet (ex. ett litet neuralt nät) justerar sig så att utgångsneuronen motsvarande rätt siffra aktiveras högst. Efter träning kan man visa en ny bild – som inte var i datat – och modellen outputtar vilken siffra det troligen är.*

**Källor:** Svenska Wikipedia: *"Förstärkningsinlärning är en av tre paradigmer tillsammans med vägledd inlärning (supervised) och ickevägledd (unsupervised)"* [102]. Den definierar supervised: *"väglett lärande: inlärning av mappning mellan indata och annoterad utdata i syfte att sedan förutsäga utdata för ny indata… ex: skräppostfilter med epostmeddelanden märkta spam/icke spam."* [170]. SAS: *"den här typen av inlärningsalgoritm (supervised) kan användas med klassificering, regression…"*. Summan: övervakat lärande = *modell lär från märkt data att generalisera till omärkt data.*

---


[1] Artificial intelligence in education | UNESCO
https://www.unesco.org/en/digital-education/artificial-intelligence

[2] [3] [4] What are AI Agents?- Agents in Artificial Intelligence Explained - AWS
https://aws.amazon.com/what-is/ai-agents/

[5] [6] [155] What Are AI Agents?
https://www.oracle.com/artificial-intelligence/ai-agents/

[7] [8] [74] What is an AI assistant? |Definition from TechTarget
https://www.techtarget.com/searchcustomerexperience/definition/virtual-assistant-AI-assistant

[9] [10] AI literacy - Wikipedia
https://en.wikipedia.org/wiki/AI_literacy

11 12 'Partnership on AI' formed by Google, Facebook, Amazon, IBM and Microsoft | Artificial intelligence (AI) | The Guardian
https://www.theguardian.com/technology/2016/sep/28/google-facebook-amazon-ibm-microsoft-partnership-on-ai-tech-firms

13 OpenAI - Wikipedia
https://en.wikipedia.org/wiki/OpenAI

14 15 156 Så genomskådar du AI-krängarnas lockrop | Computer Sweden
https://computersweden.se/article/1273551/sa-genomskadar-du-ai-krangarnas-lockrop.html

16 Exploring model welfare \ Anthropic
https://www.anthropic.com/research/exploring-model-welfare

17 Anthropic utforskar rättigheter för 'medvetna' AI-system - Neuron Expert
https://neuron.expert/news/anthropic-explores-rights-for-conscious-ai-systems/12613/sv/

18 UK Supreme Court has final say on Dabus as named inventor
https://www.juve-patent.com/cases/uk-supreme-court-dabus-named-inventor-patent-stephen-thaler/

19 UK Supreme Court (finally) issues decision in landmark 'DABUS' AI ...
https://www.lexology.com/library/detail.aspx?g=d0dd6884-a402-4c2a-8be7-cc4480a7e9d5

20 Recent Trends in Generative Artificial Intelligence Litigation in the ...
https://www.klgates.com/Recent-Trends-in-Generative-Artificial-Intelligence-Litigation-in-the-United-States-9-5-2023

21 22 23 24 AI on trial: How courts are litigating the GenAI boom
https://www.thomsonreuters.com/en-us/posts/ai-in-courts/courts-genai-boom/

25 26 27 28 AI safety - Wikipedia
https://en.wikipedia.org/wiki/AI_safety

29 30 52 89 90 Explainable AI: Så gör XAI AI-beslut begripliga och transparenta — Skräddarsydda AI-kurser & Utbildningar | AI-kurser i Stockholm & Göteborg | AIUC
https://www.aiuc.se/ai-insikter/explainable-ai-xai

31 [Solved] 6 Frklara begreppen Algoritm Vxling Tom tallinje - Studocu
https://www.studocu.com/sv/messages/question/5107582/6-forklara-begreppen-algoritm-vaxling-tom-tallinje-kompensationsberakning-fast

32 Bekanta dig med fenomenet: Algoritmer - Mediataitokoulu
https://mediataitokoulu.fi/sv/materiaalipankki/bekanta-dig-med-fenomenet-algoritmer/

33 34 71 72 78 159 What Is Algorithmic Bias? | IBM
https://www.ibm.com/think/topics/algorithmic-bias

35 [PDF] Träffsäker rekrytering i AI-eran - Lunds universitet
https://lup.lub.lu.se/student-papers/record/9156295/file/9156366.pdf

36 37 118 119 120 121 122 Interaction Bias - MDS Wiki
https://wiki.marketdomination.solutions/index.php/Interaction_Bias

38 39 40 41 42 43 44 45 46 47 85 157 162 163 164 165 Artificiell intelligens – Wikipedia
https://sv.wikipedia.org/wiki/Artificiell_intelligens

48 49 50 Artificiella neurala nätverk hjälper forskare att se in i våra skallar - Världen Om
https://varldenom.com/vetenskap/artificiella-neurala-natverk/

51 Neurala språkmodeller och transformers - Flashcards World
https://flashcards.world/flashcards/sets/bd2970b4-5b8e-4ada-aec9-09083ed5647f/

53 [PDF] AI Safety and Automation Bias | CSET
https://cset.georgetown.edu/wp-content/uploads/CSET-AI-Safety-and-Automation-Bias.pdf

54 Automation Bias - UX Method Cards
https://www.uxmethod.cards/methods/automation-bias/

55 BrainFacts.org on X: "Automation bias is the tendency to be less ...
https://twitter.com/Brain_Facts_org/status/1834693798397456646

56 57 58 59 60 AI inom bedömning väcker förhoppningar och farhågor - Skolverket
https://www.skolverket.se/skolutveckling/forskning-och-utvarderingar/artiklar-om-forskning/ai-inom-bedomning-vacker-forhoppningar-och-farhagor

61 What Are Autonomous AI Agents: Types, Benefits, and Uses | Lindy
https://www.lindy.ai/blog/autonomous-ai-agents

62 63 65 What are autonomous AI agents and which vendors offer them? | Definition from TechTarget.com
https://www.techtarget.com/searchenterpriseai/definition/autonomous-AI-agents

64 111 What Are Foundation Models? - NVIDIA Blog
https://blogs.nvidia.com/blog/what-are-foundation-models/

66 Bayesian Networks - Lightcast Skills Library
https://lightcast.io/open-skills/skills/KS120Y66FWYF6MQBR6FK

67 What is Bayesian Networks? - Dremio
https://www.dremio.com/wiki/bayesian-networks/

68 [PDF] Sannolikhetsteori i domstolen? - Lund University Publications
https://lup.lub.lu.se/student-papers/record/9069005/file/9073120.pdf

69 Beslutsträd - kategorisera ditt data och förstå varför - Multisoft
https://www.multisoft.se/kunskapsbank/beslutstrad-kategorisera-data/

70 154 Så fungerar maskininlärning och djupinlärning - Computer Sweden
https://computersweden.se/article/1283939/sa-fungerar-maskininlarning-och-djupinlarning-och-det-ar-skillnaden-mellan-teknikerna.html

73 Chatbot – Vad det är & hur det används - Pigment Webbyrå
https://pigment.se/ordlista/chatbot/

75 Best Practices for Data Quality in AI - Datagaps DQ Monitor
https://www.datagaps.com/blog/best-practices-for-data-quality-in-ai/

76 Data Quality in AI: Challenges, Importance & Best Practices
https://research.aimultiple.com/data-quality-ai/

77 The Importance of Data Quality for AI - Bloomfire
https://bloomfire.com/blog/importance-of-ai-data-quality/

79 168 4 types of statistical bias to avoid in your analyses - DataScientest
https://datascientest.com/en/4-types-of-statistical-bias-to-avoid-in-your-analyses

80 What is Omission Bias | IGI Global Scientific Publishing
https://www.igi-global.com/dictionary/a-look-at-cognitive-biases-violating-utilitarianism/123226

81 Ethics in artificial intelligence: Discriminatory biases
http://langlois.ca/en/insights/ethics-in-artificial-intelligence-discriminatory-biases/

82 83 84 Vad är djupinlärning och hur fungerar det? | NordVPN
https://nordvpn.com/sv/blog/djupinlarning/

86 Liv Boeree: On Competition, Moloch Traps, and the A.I. Arms Race
https://erictopol.substack.com/p/liv-boeree-on-competition-moloch

87 Moloch Transformed: How to Thread the Needle for a Positive AI ...
https://tamhunt.medium.com/moloch-transformed-b62ddf7f4ba9

88 Breaking Moloch: the game theory monster in AI, economics and ...
https://www.fintechfutures.com/fintech/breaking-moloch-the-game-theory-monster-in-ai-economics-and-beyond

91 Federated Learning: Decentralized and Privacy-Preserving Model ...
https://www.lyzr.ai/glossaries/federated-learning/

92 Federated Learning - an overview | ScienceDirect Topics
https://www.sciencedirect.com/topics/computer-science/federated-learning

93 What is Transfer Learning? - AWS
https://aws.amazon.com/what-is/transfer-learning/

94 What Is Few-Shot Learning? | IBM
https://www.ibm.com/think/topics/few-shot-learning

95 What Is Few Shot Learning? (Definition, Applications) | Built In
https://builtin.com/machine-learning/few-shot-learning

96 166 What is Transfer Learning? - GeeksforGeeks
https://www.geeksforgeeks.org/ml-introduction-to-transfer-learning/

97 What Are The Key Drivers Of Algorithmic Bias In Ai? → Question
https://sustainability-directory.com/question/what-are-the-key-drivers-of-algorithmic-bias-in-ai/

98 99 [PDF] BIAS REINFORCEMENT IN LARGE LANGUAGE MODELS
https://openreview.net/attachment?id=c5bjw7hqix&name=pdf

100 [PDF] Measuring the Concentration Reinforcement Bias of Recommender ...
https://ceur-ws.org/Vol-1441/recsys2015_poster11.pdf

101 Bias from a philosophical standpoint - OpenAI Developer Community
https://community.openai.com/t/bias-from-a-philosophical-standpoint/606050

102 103 104 105 Förstärkningsinlärning – Wikipedia
https://sv.wikipedia.org/wiki/F%C3%B6rst%C3%A4rkningsinl%C3%A4rning

106 123 132 133 134 135 136 151 170 Maskininlärning – Wikipedia
https://sv.wikipedia.org/wiki/Maskininl%C3%A4rning

107 108 109 Generativ AI i akademiskt skrivande - Skrivguiden.se
https://skrivguiden.se/skriva/generativ-ai-i-akademiskt-skrivande/

110 What Are Foundation Models? - IBM
https://www.ibm.com/think/topics/foundation-models

112 What Are Foundation Models? | Built In
https://builtin.com/artificial-intelligence/foundation-models

113 [PDF] Hybrid AI - Plattform Lernende Systeme
https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen_EN/KI_Kompakt_EN/PLS_AI_ataglance_Hybrid_AI.pdf

114 A vision on Hybrid AI for military applications
https://resolver.tno.nl/uuid:e36391b2-7580-4b93-8247-8815821896d0

115 Converging hybrid AI and enterprise modeling - metaphacts Blog
https://blog.metaphacts.com/the-future-of-information-systems-converging-hybrid-ai-and-enterprise-modeling

116 Intelligent tutoring system - Wikipedia
https://en.wikipedia.org/wiki/Intelligent_tutoring_system

117 Intelligent Tutoring System | Encyclopedia MDPI
https://encyclopedia.pub/entry/28717

124 126 Vad är skillnaden mellan regression och klassificering i ...
https://sv.eitca.org/artificial-intelligence/eitc-ai-tff-tensorflow-fundamentals/tensorflow-in-google-colaboratory/using-tensorflow-to-solve-regression-problems/examination-review-using-tensorflow-to-solve-regression-problems/what-is-the-difference-between-regression-and-classification-in-machine-learning/

125 Beslutsträd Flashcards | Quizlet
https://quizlet.com/se/960132990/beslutstrad-flash-cards/

127 128 158 Kunskapsrepresentation och resonerande – AI Competence for Sweden
https://ai-competence.se/topic/kunskapsrepresentation-och-resonerande/

129 Learning analytics - Wikipedia
https://sv.wikipedia.org/wiki/Learning_analytics

130 Varför är Learning Analytics så viktigt? - SEAtS Software
https://www.seatssoftware.com/sv/the-importance-of-learning-analytics/

131 Analys av inlärning kan förnya pedagogiken - Universitetsläraren
https://universitetslararen.se/2016/11/24/analys-av-inlarning-kan-fornya-pedagogiken/

137 What Is Meta Learning? - IBM
https://www.ibm.com/think/topics/meta-learning

138 Meta Learning: How Machines Learn to Learn - DataCamp
https://www.datacamp.com/blog/meta-learning

139 AI, Moloch, and the race to the bottom - IAI TV
https://iai.tv/articles/ai-moloch-and-the-race-to-the-bottom-auid-2509

140 Moloch - AI Alignment Forum
https://www.alignmentforum.org/w/moloch

141 The Rise of Multimodal AI: Combining Text, Image, and Audio ...
https://dev.to/hakeem/the-rise-of-multimodal-ai-combining-text-image-and-audio-understanding-32m3

142 Multimodal AI | Google Cloud
https://cloud.google.com/use-cases/multimodal-ai

143 Multimodal AI: A Guide to Open-Source Vision Language Models
https://www.bentoml.com/blog/multimodal-ai-a-guide-to-open-source-vision-language-models

144 Multimodal AI: Bridging Text and Visual Data - Medium
https://medium.com/@API4AI/multimodal-ai-bridging-text-and-visual-data-60f181138914

145 ELI5: Vad är Naturlig Språkbehandling : r/explainlikeimfive - Reddit
https://www.reddit.com/r/explainlikeimfive/comments/axnk0o/eli5_what_is_natural_language_processing/?tl=sv

146 Datateknik, avancerad nivå, Naturlig språkbehandling, 3 hp - Örebro ...
https://www.oru.se/utbildning/kurser/kurs/datateknik-avancerad-niva-naturlig-sprakbehandling-dt712a

147 Handledning för naturlig språkbehandling: Vad är NLP? Exempel
https://www.guru99.com/sv/nlp-tutorial.html

148 Djupinlärning jämfört med maskininlärning - Azure Machine Learning
https://learn.microsoft.com/sv-se/azure/machine-learning/concept-deep-learning-vs-machine-learning?view=azureml-api-2

149 Object Detection vs Object Recognition vs Image Segmentation
https://www.geeksforgeeks.org/object-detection-vs-object-recognition-vs-image-segmentation/

150 What is Object Recognition? - Roboflow Blog
https://blog.roboflow.com/what-is-object-recognition/

152 Maskininlärning - vad är det och varför är det viktigt? - SAS
https://www.sas.com/sv_se/insights/analytics/machine-learning.html

153 Proactive Data Quality Management: A Key to Successful AI Adoption!
https://medium.com/@sahin.samia/proactive-data-quality-management-a-key-to-successful-ai-adoption-16f229d8a881

160 [PDF] Self-supervised learning on tabular data - Lund University Publications
https://lup.lub.lu.se/student-papers/record/9158488/file/9158495.pdf

161 data2vec: En milstolpe i självövervakat lärande - Unite.AI
https://www.unite.ai/sv/data2vec/

167 SHOW-O: En enda transformator som förenar multimodal ... - Unite.AI
https://www.unite.ai/sv/visa-en-enda-transformator-som-f%C3%B6renar-multimodal-f%C3%B6rst%C3%A5else-och-generering/

169 A Summary of Approaches to Few-Shot Learning - arXiv
https://arxiv.org/abs/2203.04291

ChatGPT

# Ordlista – fyra centrala begrepp inom artificiella neurala nätverk

## Input

### Definition

**Input** (indata) syftar på de data eller signaler som matas in i ett neuralt nätverk för bearbetning. Det kan vara siffror, text, bilder, ljud eller andra former av råinformation som utgör utgångspunkten för nätverkets beräkningar.

### Beskrivning

I ett artificiellt neuralt nätverk utgör input själva startpunkten för informationsflödet – det vill säga allt det material nätverket får att arbeta med. Input presenteras ofta genom ett **input-lager** (inmatningslager), där varje nod i detta lager motsvarar en del av indata (t.ex. en pixel i en bild eller ett värde i en datauppsättning). Input-lagrets uppgift är att ta emot den råa informationen och förmedla den vidare in i nätverket [1]. Själva *bearbetningen* sker sedan i efterföljande lager av nätverket, men utan en korrekt och relevant input kommer nätverket inte att kunna prestera väl.

Kvaliteten och utformningen på input är avgörande för hur bra nätverket kan lära sig och lösa sin uppgift. Det finns ett klassiskt talesätt inom dataanalys: *"skräp in, skräp ut"*, vilket betyder att om ett AI-system tränas på bristfällig eller missvisande input-data riskerar även output (resultatet) att bli bristfälligt. I praktiken lägger man därför stor vikt vid att samla in rätt sorts data och ofta förbehandla (städa, normalisera eller strukturera) dessa input-data innan de skickas in i nätverket. På så sätt ökar man chansen att nätverket hittar de mönster som är relevanta för problemet man vill lösa.

### Exempel

*En AI-baserad matteträningsapp* för elever kan illustrera begreppet input. Antag att en elev matar in svaret "42" på en matematikuppgift i appen. Denna siffra (tillsammans med information om uppgiften) utgör input till det neurala nätverk som driver appens AI-funktion. Nätverket tar emot elevens svar som indata och bearbetar det sedan genom sina lager för att avgöra om svaret är korrekt. I detta fall är alltså elevens inmatade tal **42** input – det är utifrån den informationen som AI-modellen arbetar för att ge feedback. Om input förändras (t.ex. om eleven istället matar in "43") så får nätverket en annan utgångspunkt, vilket kan leda till en annan **output** (t.ex. feedback om att svaret är fel och en ledtråd till hur man räknar ut uppgiften).

### Källor


- IBM Cloud Education. (2021, 6 oktober). *What is a neural network?* (IBM-artikel) – Beskriver att varje neuralt nätverk består av ett **input-lager** som tar emot data, eventuella dolda lager och ett **output-lager** som ger resultat [2].
- NordVPN. (2023, 22 maj). *Vad är neurala nätverk och hur fungerar de?* [Blogginlägg] – Förklarar att **input-lagret** i ett neuralt nätverk helt enkelt är det lager som tar emot information och matar in

det i nätverket [1]. Diskuterar även hur data går från input-lagret, genom dolda lager, till output-lagret.

# Output

## Definition

**Output** (utdata) är det resultat som ett neuralt nätverk producerar efter att ha bearbetat sin input. Det kan vara i form av en kategori, ett numeriskt värde, en textsträng, en signal eller annan typ av svar som nätverket levererar som sitt **utkommande** besked på given indata.

## Beskrivning

Output representerar nätverkets slutgiltiga svar eller beslut baserat på den inmatade datan. I arkitekturen för ett neuralt nätverk är det **output-lagret** (utmatningslagret) som samlar upp informationen från de föregående lagren och generar nätverkets *slutresultat*. Beroende på vilken uppgift AI-modellen har kan output ta olika former. För en bildigenkänningsmodell kan output vara en klassificering (t.ex. "Det här är en katt" eller "Det här är en hund" med viss sannolikhet). För en regressionsmodell kan output vara ett numeriskt värde (t.ex. det prognostiserade priset på en vara). I en språkmodell kan output vara en genererad mening eller text. Gemensamt är att output är det vi människor eller externa system tar emot från AI:n – det vi bryr oss om i praktiken.

Inom maskininlärning utvärderas output ofta mot önskat facit. Under träning av ett neuralt nätverk jämförs nätverkets output med det förväntade rätta svaret (om sådant finns) för att räkna ut ett felvärde, vilket sedan används för att justera nätverkets interna parametrar. På så vis lär sig nätverket att förbättra sina framtida output. I en användningsmiljö, särskilt inom utbildning, kan output från ett pedagogiskt AI-system anta formen av feedback eller beslut som direkt påverkar elevens lärandeupplevelse. Till exempel kan en output vara *"Rätt svar!"* alternativt *"Försök igen, tänk på att använda Pythagoras sats"* – alltså konkret respons som eleven möts av. En pålitlig och pedagogiskt utformad output är viktig för att användaren (eleven eller läraren) ska ha nytta av AI-systemet.

## Exempel

*En språkinlärningsapp med AI* kan belysa output-begreppet tydligt. Föreställ dig en app där en elev övar på engelskt uttal genom att tala en mening i mikrofonen. Det neurala nätverket i appen analyserar elevens inspelade röst (input). **Output** från nätverket skulle i detta fall kunna vara en återkoppling på uttalet. Till exempel kan AI-systemet generera meddelandet: *"Bra jobbat! Din satsmelodi är korrekt, men öva gärna på att uttala 'R' mer distinkt."* Denna text är nätverkets output – det vill säga, resultatet av AI:ns bearbetning av elevens tal. Outputen levereras direkt till eleven som feedback i appen. Om eleven istället hade uttalat meningen annorlunda (annan input) skulle också output-meddelandet kunna blivit annorlunda, anpassat efter det specifika uttalet. På så vis representerar output det svar eller den åtgärd som AI:t ger tillbaka till användaren baserat på det den fått som indata.

## Källor

- AI Fakta. (2025). *Neurala nätverk* – Förklarar att ett **utmatningslager (output layer)** ger nätverkets förutsägelser eller beslut, och att antalet noder i output-lagret beror på uppgiften (t.ex. en nod för binär ja/nej-klassificering eller flera noder för multi-klass problem) [3].
- IBM Cloud Education. (2021, 6 oktober). *What is a neural network?* – Betonar att ett neuralt nätverk avslutas med ett **output layer** som producerar slutresultatet av nätverkets beräkningar

[2]. Denna output kan vara allt från en enkel signal till komplex information, beroende på nätverkets design och syfte.

# Nod

## Definition

**Nod** (även kallad artificiell neuron) är den grundläggande beräkningsenheten i ett neuralt nätverk. En nod tar emot en eller flera input-signaler, utför en beräkning (t.ex. summering av signalerna plus en aktiveringsfunktion) och skickar sedan ut en output-signal till andra noder eller som bidrag till nätverkets slutgiltiga resultat.

## Beskrivning

En nod i ett neuralt nätverk kan liknas vid en kraftigt förenklad modell av en biologisk neuron i hjärnan. Precis som neuronen tar emot signaler via sina dendriter, tar den artificiella noden emot numeriska värden som input. Varje inkommande signal vägs med en viss **vikt** – en parameter som indikerar signalens relativa betydelse – och noden summerar dessa viktade inputvärden. Därefter appliceras en **aktiveringsfunktion** (en matematisk funktion, ofta icke-linjär) på summan för att avgöra nodens utgående aktivering. Resultatet blir nodens **output**, som kan ses som nodens *”svar”* givet de input den fick. Om output-värdet passerar ett visst tröskelvärde (beroende på aktiveringsfunktionen) så säger man att noden "eldar av" eller aktiveras – den skickar då sin signal vidare. Detta motsvarar analogt att en biologisk neuron skickar iväg en nervimpuls om den totala stimuleringen är tillräcklig.

Noderna är typiskt organiserade i lager, där output från noder i ett lager kopplas till input för noder i nästa lager [4]. Varje nod bidrar alltså med sin beräkning till helheten, och det är genom samverkan av många noder som neurala nätverk kan utföra komplexa uppgifter. Under **träning** av nätverket justeras vikterna på varje nods ingångar gradvis, så att varje nod lär sig att reagera på rätt sätt på olika mönster i datan. Antalet noder och hur de kopplas påverkar nätverkets kapacitet: fler noder (och därmed fler parametrar/vikter) kan representera mer komplexa samband, men kräver också mer data och beräkningskraft för att tränas ordentligt. Noderna utgör därför “byggstenarna” i det neurala nätverket – de utför de grundoperationer som, när de staplas och kombineras i stor skala, leder till nätverkets intelligenta beteende.

## Exempel

*Handskriftstolkning med neurala nätverk* kan exemplifiera vad en enskild nod gör. Tänk dig en AI som ska känna igen handskrivna siffror (0–9) från bilder – något som kan användas i skolan för att automatiskt läsa av inscannade prov. När en elev har skrivit siffran "7" på papper och detta skannas in som en bild, kommer nätverkets första lager av noder ta emot pixlarna som input. En **nod** i ett av de dolda lagren kan då ha lärt sig att reagera på ett visst mönster i bilden. Till exempel kanske en specifik nod aktiveras starkt av horisontella linjer. När bilden av "7" matas genom nätverket kommer just den noden att få hög output, eftersom en handskriven 7:a innehåller en tydlig horisontell streck upptill. Denna nod “tänder” alltså på det draget och skickar vidare sin signal till nästa lager. Andra noder kan under tiden reagera på diagonala linjer eller vinklar i samma bild. I nästa lager kombineras signalerna från många sådana noder, vilket hjälper nätverket att avgöra att bilden föreställer just siffran 7. I detta exempel illustrerar en enskild nods output (detektionen av en horisontell linje) hur noder samarbetar: varje nod bidrar med att känna igen ett delmönster, och tillsammans möjliggör de att det neurala nätverket kan tolka en komplex input (hela bilden av siffran).

### Källor

- Wikipedia. (2023). *Artificiellt neuronnät* – Slår fast att neuronen (nod, enhet) är grundbyggstenen i ett neuralt nätverk [4]. Varje artificiell neuron tar emot flera ingångar och har en utgång, ungefär som en biologisk neuron har dendriter och ett axon.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press – Kapitel 6 förklarar flerskiktsperceptroner och artificiella neuroner matematiskt. Boken beskriver hur varje **nod** beräknar en viktad summa av sina indata och applicerar en aktiveringsfunktion för att producera sin **output**, vilket utgör grunden för hur neurala nätverk bygger upp komplexa funktioner.

## Lager

### Definition

**Lager** i ett neuralt nätverk syftar på ett skikt av noder som verkar på samma nivå i nätverkets struktur. Ett typiskt neuralt nätverk består av ett **input-lager** (som tar emot indata), ett eller flera **dolda lager** (som bearbetar informationen på olika abstraktionsnivåer) samt ett **output-lager** (som levererar slutresultatet). Varje lager tar emot signaler från föregående lager och skickar vidare sina output-signaler till nästa lager i kedjan.

### Beskrivning

Att neurala nätverk är uppbyggda av lager innebär att beräkningen sker stegvis i flera nivåer. Först tar input-lagret in rådata och distribuerar den till nätverkets första dolda lager. **Dolda lager** (också kallade *hidden layers* eftersom de inte direkt observeras utifrån) utför majoriteten av beräkningarna – de omvandlar input-data till mer och mer användbara representationer. Varje lager tar emot *aktiveringar* (signaler) från lagret innan, utför sina beräkningar genom noderna, och skickar sedan vidare nya aktiveringar till nästa lager. Slutligen samlas all bearbetning i **output-lagret**, som producerar nätverkets svar. Genom denna lager-på-lager-struktur kan nätverket successivt extrahera mer avancerade *drag* eller mönster ur datan [1] [5]. Ett enkelt neuralt nätverk har minst tre lager (input, ett dolt, output), medan *djupa neurala nätverk* kan ha många dolda lager – det är just antalet lager som åsyftas med "djup" i **djupinlärning** (eng. *deep learning*). Fler lager gör det möjligt för nätverket att lära sig mer komplexa funktioner, eftersom varje lager kan bygga vidare på de mönster som föregående lager identifierat.

Lagerstrukturen är central för varför neurala nätverk är så kraftfulla. **Hierarkisk abstraktion** är en följd av flera lager: tidiga lager kan lära sig känna igen mycket enkla inslag i datan, medan senare lager sammanför dessa inslag till något mer meningsfullt. Till exempel, i ett bildigenkänningsnätverk kan första lagret leta efter enkla linjer eller färgövergångar, nästa lager kombinera linjer till grundläggande former, och ännu senare lager identifiera komplexa objekt som ansikten eller bokstäver. Varje lager höjer abstraktionsnivån [6]. Denna flerstegsprocess gör att nätverket slutligen kan förstå högnivåegenskaper i datan – något enskilda lager inte klarar av på egen hand. Att öka antalet dolda lager (och noder) ökar ofta nätverkets förmåga att lösa svåra problem, men det kan också göra träningen mer krävande och öka risken för överanpassning om man inte har tillräckligt med data. Forskning inom djupinlärning har dock visat att väldigt djupa nätverk, rätt tränade, kan upptäcka intrikata mönster i stora datamängder och nå imponerande prestanda inom allt från bild- och taligenkänning till spelstrategier [6]. Sammanfattningsvis utgör lagren “nivåerna” i nätverkets tänkande, där varje nivå förädlar informationen ytterligare – en princip som är grundläggande för modern AI.

## Exempel

*AI-baserad textbedömning* kan ge ett pedagogiskt exempel på hur flera lager samverkar. Föreställ dig ett neuralt nätverk som används för att ge feedback på elevtexter (såsom essäer eller uppsatser). **Input-lagret** tar först emot elevens text, t.ex. genom att varje ord eller varje bokstav kodas till siffror som nätverket kan hantera. Därefter skickas dessa kodade ord vidare in i nätverkets första **dolda lager**. I detta lager kanske nätverket börjar med att identifiera grundläggande språkliga mönster – exempelvis kan vissa noder här reagera på stavningsfel eller enklare grammatiska avvikelser. Signalerna från det första lagret förs sedan vidare till nästa **lager**, där analysen fördjupas. Här skulle nätverket kunna sammanfatta textens innehåll eller strukturera meningarna för att förstå textens uppbyggnad: vissa noder kan aktiveras av en röd tråd i argumentationen, andra av förekomsten av nyckelord som indikerar textens tema. Om nätverket är djupare kan ytterligare lager upptäcka ännu mer abstrakta drag, till exempel textens tonfall eller kreativitet, genom att kombinera informationen från tidigare lager. Slutligen når bearbetningen **output-lagret**, där all denna insamlade information vägs samman. Output-lagret kan då ge sitt utlåtande i form av en *återkoppling* till eleven. Det kan röra sig om ett betygsförslag (t.ex. "B") och/eller konkreta kommentarer såsom *"Dispositionen är tydlig och språket varierat, men resonemangen kan utvecklas mer."* Denna slutliga feedback representerar nätverkets samlade tolkning av texten. Genom att informationen fått passera genom flera lager – från rå text till lingvistiska detaljer, vidare till innehållsanalys och slutligen bedömning – kan AI-systemet leverera en nyanserad och användbar output. Exemplet illustrerar hur varje lager i nätverket bidrar med ett nytt perspektiv på datan, vilket möjliggör en helhetsförståelse av elevens text som vore svår att uppnå med en enklare, endimensionell analys.

## Källor


- NordVPN. (2023, 22 maj). *Vad är neurala nätverk och hur fungerar de?* – Förklarar nätverksarkitekturen med **minst tre lager: input-lager, dolt lager och output-lager**, samt hur data flödar framåt genom lagren [1] [5]. Artikeln noterar även att om problemet är mer komplext kan flera dolda lager användas för djupare analys.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning*. *Nature, 521*(7553), 436–444 – En översiktsartikel som etablerar grunden för djupinlärning. Författarna förklarar att djupinlärning använder “flera bearbetnings**lager** för att lära sig representationer av data på flera abstraktionsnivåer” [6]. Detta belyser hur varje ytterligare lager i ett neuralt nätverk möjliggör en högre nivå av förståelse (t.ex. från enkla mönster till komplexa koncept). Artikeln exemplifierar också hur djupa nät med många lager dramatiskt förbättrat prestandan i bild- och taligenkänning.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press – Standardlitteratur inom djupinlärning. Kapitlen om **flerskiktsnätverk** (multilayer networks) förklarar hur olika lager – inklusive **dolda lager** – samarbetar och hur träningsalgoritmer såsom backpropagation sprider fel genom lagren för att justera vikter. Boken understryker att det är just användningen av många på varandra följande beräkningslager som ger moderna neurala nätverk deras enorma uttryckskraft och förmåga att modellera komplexa samband [6]. (Tillgänglig online: *deeplearningbook.org*)


---


[1] [5] Vad är neurala nätverk och hur fungerar de? | NordVPN
https://nordvpn.com/sv/blog/neurala-natverk/

[2] What is a Neural Network? | IBM
https://www.ibm.com/think/topics/neural-networks

3 Neurala nätverk - AI Fakta - allt du behöver veta om Artificiell Intelligens
https://aifakta.se/grunderna-inom-ai/neurala-natverk/

4 Artificiellt neuronnät – Wikipedia
https://sv.wikipedia.org/wiki/Artificiellt_neuronn%C3%A4t

6 (PDF) Deep Learning
https://www.researchgate.net/publication/277411157_Deep_Learning