

OBEROENDE

EXPERTGRUPP PÅ HÖG NIVÅ FÖR AI-FRÅGOR

INRÄTTAD AV EUROPEISKA KOMMISSIONEN I JUNI 2018



EN DEFINITION AV AI: VIKTIGASTE FÖRMÅGOR OCH DISCIPLINER

Definition framtagen för att användas i gruppens leverabler

En definition av AI: Viktigaste förmågor och vetenskapliga discipliner

Kommissionens expertgrupp på hög nivå för AI-frågor

Ansvarsfriskrivning och användning av detta dokument: Följande beskrivning och definition av AI-förmågor och forskningsområden är en mycket grov förenkling av den senaste tekniken. Syftet med detta dokument är inte att exakt och uttömmande definiera alla AI-tekniker och AI-förmågor, utan att kortfattat beskriva den gemensamma uppfattning om denna disciplin som expertgruppen på hög nivå använder i sina leverabler. Vi hoppas dock att detta dokument också kan användas som en bra utgångspunkt för att utbilda människor som inte är experter på AI och som sedan kan följa upp detta med en mer omfattande och djupgående reflektion över AI, för att få mer exakta kunskaper om denna disciplin och teknik.

AI HLEG är en oberoende expertgrupp som inrättades av Europeiska kommissionen i juni 2018.

Kontakt Nathalie Smuha – Samordnare AI HLEG
E-post CNECT-HLG-AI@ec.europa.eu

Europeiska kommissionen
B-1049 Bryssel

Dokumentet offentliggjordes den X april 2019.

Ett första utkast till detta dokument offentliggjordes den 18 december 2018, tillsammans med det första förslaget till expertgruppens etiska riktlinjer för tillförlitlig AI. Det reviderades med hänsyn till de synpunkter som inkom via den europeiska AI-alliansen och det öppna samråd om förslaget till riktlinjer. Vi vill uttryckligen och varmt tacka alla som bidragit med återkoppling på det första utkastet till dokumentet.

Varken Europeiska kommissionen eller någon person som agerar för kommissionens räkning ansvarar för hur nedanstående information kan komma att användas. Expertgruppen på hög nivå för AI-frågor ansvarar ensam för innehållet i detta arbetsdokument. Personal vid kommissionen har bidragit till utarbetandet av detta dokument, men de synpunkter som framförs här återspeglar expertgruppens ståndpunkter och kan inte under några omständigheter anses vara ett uttryck för kommissionens officiella ställningstagande.

Mer information om expertgruppen på hög nivå för AI-frågor finns här: (<https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence>).

Vidareutnyttjande av kommissionens handlingar regleras i beslut 2011/833/EU (EUT L 330, 14.12.2011, s. 39). Tillstånd för användning eller reproduktion av bilder och annat material som inte omfattas av EU:s upphovsrätt ska sökas direkt av upphovsmannen.

EN DEFINITION AV AI:

VIKTIGASTE FÖRMÅGOR OCH VETENSKAPLIGA DISCIPLINER

Vi utgår från följande definition av artificiell intelligens (AI), som anges i kommissionens meddelande om AI¹:

”Artificiell intelligens avser system som uppvisar intelligent beteende genom att analysera sin miljö och vidta åtgärder – med viss grad av självständighet – för att uppnå särskilda mål.

AI-baserade system kan vara helt programvarubaserade och fungera i den virtuella världen (t.ex. röstassistenter, bildanalysprogram, sökmotorer, tal- och ansiktsigenkänningssystem), eller inbäddas i hårdvaruenheter (t.ex. avancerade robotar, självkörande bilar, drönare eller tillämpningar för sakernas internet).”

I detta dokument utvecklar vi definitionen för att förtydliga vissa aspekter av AI som vetenskaplig disciplin och som teknik, i syfte att undvika missförstånd, skapa en delad gemensam kunskap om AI som kan användas på ett fruktbart sätt även av dem som inte är experter på AI, samt lämna användbara uppgifter som kan användas i diskussionen om både de etiska riktlinjerna för AI och de politiska rekommendationerna om AI.

1. AI-system

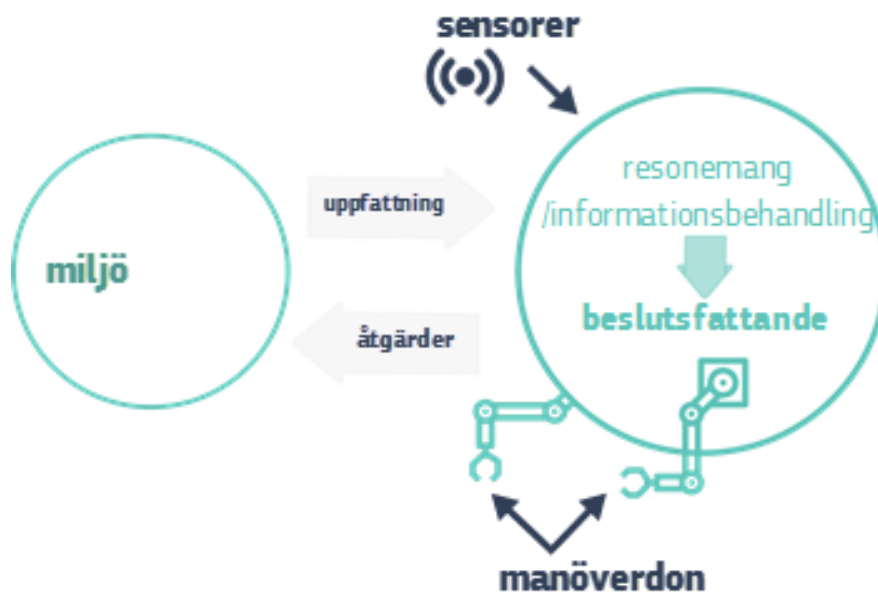
Begreppet AI innehåller en uttrycklig hänvisning till begreppet intelligens. Eftersom intelligens (hos både maskiner och människor) emellertid är ett vagt begrepp, trots att det har studerats ingående av psykologer, biologer och neurologer, använder AI-forskare oftast begreppet rationalitet. Med det avses förmågan att välja den bästa åtgärden att vidta för att uppnå ett visst mål, med vissa kriterier som ska optimeras och med hänsyn till tillgängliga resurser. Rationalitet är förstås inte den enda beståndsdelen i begreppet intelligens, men den är en betydande del i det.

I fortsättningen kommer vi att använda begreppet *AI-system* i betydelsen alla AI-baserade komponenter, programvaror och/eller maskinvaror. Ofta är AI-system *inbäddade* som komponenter i större system, snarare än fristående system.

Ett AI-system är alltså först och främst rationellt, enligt en av de mest använda läroböckerna i AI². Men hur uppnår ett AI-system rationalitet? Som anges i första meningen i den ovannämnda arbetsdefinitionen av AI gör den det genom att uppfatta den miljö i vilken systemet befinner sig genom några sensorer och på så sätt samla in och tolka data, resonera om det som uppfattas, eller bearbeta den information som härleds ur dessa data, och avgöra vilken den bästa åtgärden är för att sedan agera utifrån det, genom några manöverdon, för att på så sätt eventuellt modifiera miljön. AI-system kan använda symboliska regler eller lära sig en numerisk modell. De kan också anpassa sitt beteende genom att analysera hur miljön har påverkats av deras föregående åtgärder. Illustrationen av ett AI-system i figur 1 kan underlätta.

¹ Kommissionens meddelande till Europaparlamentet, rådet, Europeiska ekonomiska och sociala kommittén och Regionkommittén – *Artificiell intelligens för Europa*, Bryssel, 25.4.2018, COM(2018) 237 final.

² ”Artificial Intelligence: A Modern Approach”, S. Russell och P. Norvig, Prentice Hall, 3:e utgåvan, 2009.



Figur 1: En schematisk avbildning av ett AI-system.

Sensorer och perception. I figur 1 avbildas systemets sensorer som en wifi-symbol. I verkligheten kan de vara kameror, mikrofoner, ett tangentbord, en webbplats eller andra inmatningsenheter, och även sensorer för fysiska kvantiteter (t.ex. temperatur, tryck, avstånd, kraft/vridmoment, taktila sensorer). Generellt behöver vi ge AI-systemet sensorer som är lämpliga för att uppfatta de data som finns i miljön och som är relevanta för det mål som AI-systemet som den mänskliga designern har satt upp för det. Om vi t.ex. vill bygga ett AI-system som automatiskt rengör golvet i ett rum när det är smutsigt, skulle sensorerna bl.a. kunna vara kameror som tar bilder av golvet.

När det gäller insamlade data är det ofta användbart att göra åtskillnad mellan strukturerade och ostrukturerade data. *Strukturerade data* är data som är organiserade enligt på förhand fastställda modeller (t.ex. i en relationsdatabas), medan *ostrukturerade data* inte har någon känd organisation (t.ex. en bild eller ett stycke text).

Resonemang/informationsbehandling och beslutsfattande. Kärnan i ett AI-system är dess modul för resonemang/informationsbehandling, som använder de data som sensorerna förmedlar och föreslår en åtgärd som ska vidtas, med hänsyn till det mål som ska uppnås. Det betyder att de data som sensorerna samlar in måste omvandlas till information som modulen för resonemang/informationshantering kan förstå. För att fortsätta med vårt exempel på ett städnings-AI-system, kommer kameran att lämna en bild av golvet till modulen för resonemang/informationsbehandling. Nu behöver modulen avgöra om golvet ska städas eller ej (dvs. vilken är den bästa åtgärden för att uppnå det önskade målet). Det kan verka enkelt för oss människor att gå från bilden av ett golv till beslutet om huruvida det behöver rengöras, men det är inte så lätt för en maskin, eftersom en bild bara är en sekvens av nollor och ettor. Modulen för resonemang/informationsbehandling måste alltså göra följande:

1. Tolka bilden för att avgöra om golvet är rent eller ej. Det betyder generellt att modulen måste kunna omvandla data till information och modellera den informationen på ett kortfattat sätt samtidigt som alla relevanta data måste komma med (i detta fall om golvet är rent eller ej).
2. Resonera om denna kunskap eller behandla denna information för att framställa en numerisk modell (dvs. en matematisk formel) för att avgöra vilken åtgärd som är bäst. I detta exempel betyder det att om den information som härleds ur bilden säger att golvet är smutsigt, är den bästa åtgärden att aktivera städning. I annat fall är den bästa åtgärden att inte göra något alls.

Observera att begreppet "beslut" används i vid bemärkelse, som handlingen att välja den åtgärd som ska vidtas. Det betyder inte nödvändigtvis att AI-systemen är helt självständiga. Ett beslut kan också bestå av ett val av en rekommendation som ska ges till en människa, som sedan fattar det slutliga beslutet.

Manövrering. När AI-systemet har beslutat om en åtgärd är det redo att utföra åtgärden genom de manöverdon som systemet har tillgång till. I teckningen ovan avbildas manöverdonen som ledade armar, men de behöver inte vara fysiska. Manöverdonen kan även bestå av programvara. I vårt städningsexempel skulle AI-systemet kunna producera en signal som aktiverar en dammsugare, om åtgärden är att rengöra golvet. Som ett annat exempel agerar ett konversationssystem (dvs. en chattrobot) genom att generera textmeddelanden som svar på användarens yttranden.

Den åtgärd som vidtas kommer eventuellt att modifiera miljön, vilket betyder att nästa gång behöver systemet använda sina sensorer på nytt för att uppfatta eventuell ny information från den modifierade miljön.

Rationella AI-system väljer inte alltid den bästa åtgärden för sitt mål. I sådana fall uppnår de endast *begränsad rationalitet* beroende på begränsningar i resurser såsom tid eller datorkapacitet.

Rationella AI-system är en mycket grundläggande version av AI-system. De modifierar miljön, men anpassar inte sitt beteende över tid för att uppnå sitt mål. Ett *lärande rationellt system* är ett rationellt system som efter att ha vidtagit en åtgärd utvärderar miljöns nya status (genom perception) för att avgöra hur framgångsrik åtgärden var för att sedan anpassa sina resonemangsregler och beslutsmetoder.

2. AI som vetenskaplig disciplin

Vi har i det föregående avsnittet gett en mycket enkel och abstrakt beskrivning av AI-system, genom tre viktiga förmågor: perception, resonemang/beslutsfattande samt manövrering. Detta räcker för att vi ska kunna presentera och förstå de flesta AI-tekniker och deldiscipliner som för närvarande används för att bygga AI-system, eftersom alla hänvisar till systemens olika förmågor. Generellt kan all sådan teknik delas in i två huvudgrupper beroende på förmågan till *resonemang* och *lärande*. Robotik är ett annat mycket relevant område.

Resonemang och beslutsfattande Denna grupp av tekniker omfattar kunskapsrepresentation och resonande, planering, schemaläggning sökning och optimering. Dessa tekniker möjliggör resonemang om de data som kommer från sensorerna. För att kunna göra detta behöver man omvandla data till kunskap. Ett område för AI handlar därför om hur man bäst kan modellera denna kunskap (*kunskapsrepresentation*). När kunskapen har modellerats är nästa steg att resonera om den (*kunskapsresonemang*), vilket omfattar att dra slutsatser genom symboliska regler, *planering* och *schemaläggning* av aktiviteter, *sökning* i en stor uppsättning lösningar samt *optimering* med hänsyn till alla presumtiva lösningar på ett problem. Det sista steget är att avgöra vilken åtgärd som ska vidtas. Resonemangs-/beslutsdelen i ett AI-system är oftast mycket komplicerat och kräver en kombination av flera av de tekniker vi beskriver ovan.

Inläring. Denna grupp av tekniker omfattar maskininläring, neuronät, fördjupat lärande, beslutsträd och många andra inlärningsmetoder. Med dessa tekniker kan ett AI-system lära sig att lösa problem som inte går att specificera exakt, eller vars lösningsmetod inte går att beskriva med symboliska resonemangsregler. Exempel på sådana problem är de som hänger samman med perceptionsförmåga, t.ex. *tal-* och *språkförståelse* samt *datorseende* eller *förutsägelser om beteende*. Dessa saker är till synes enkla, eftersom de faktiskt ofta är enkla för människor. Men de är inte lika enkla för AI-system, eftersom de inte kan förlita sig på sunt förnuft (åtminstone inte än) och de är extra svåra när systemet måste tolka ostrukturerade data. Det är här teknik som bygger på metoden med *maskininläring* kommer väl till pass. Maskininläringsteknik kan dock användas till mycket mer än enbart perception. Maskininläringsteknik bygger upp en numerisk modell (dvs. en matematisk formel) som används för att beräkna beslutet utifrån datan.

Maskininläring finns i flera former. De vanligaste metoderna är *övervakad inläring*, *oövervakad inläring* och *förstärkt inläring*.

I övervakad maskininlärning ger vi systemet exempel på input-output-beteende i stället för beteenderegler, i förhoppningen att systemet ska kunna dra generella slutsatser utifrån exemplen (som oftast beskriver tidigare händelser) och bete sig väl även i situationer som inte visas i exemplen (som kan uppstå i framtiden). I vårt exempel skulle vi ge systemet många exempel på bilder av ett golv med tillhörande tolkning (dvs. om golvet är rent eller ej på den bilden). Om vi ger tillräckligt många exempel som innehåller tillräckligt stor variationsrikedom och inkluderar tillräckligt många, eller de flesta, situationer, kommer systemet genom sin maskininlärningsalgoritm att kunna generalisera för att kunna tolka bilder på golv som det aldrig har sett förut. I vissa metoder för maskininlärning används algoritmer som baseras på begreppet *neuronät*, som fritt har inspirerats av den mänskliga hjärnan i den bemärkelsen att det har ett nätverk av små processenheter (motsvarande våra neuroner) med en mängd viktade kopplingar mellan sig. I ett neuronät kommer input i form av data från sensorerna (i vårt exempel bilden på golvet) och dess output är tolkningen av bilden (i vårt exempel huruvida golvet är rent eller ej). Under analysen av exemplen (nätverkets *träningsfas*) justeras kopplingarnas viktning för att i så hög grad som möjligt motsvara det som de tillgängliga exemplen visar (dvs. för att minimera avvikelserna mellan förväntad output och den output som nätverket räknar fram). I slutet av träningsfasen testas neuronätets beteende med exempel som det inte har fått se förut, för att kontrollera att träningen har gett önskat resultat.

Det är viktigt att vara medveten om att denna metod (precis som all maskininlärningsteknik) alltid har en viss felprocent, även om den oftast är liten. Ett avgörande begrepp är därför *korrekthet* – ett mått på hur stor andelen korrekta svar är.

Det finns flera slags neuronät och metoder för maskininlärning. En av de mest framgångsrika i dag är *fördjupad inlärning*. Denna metod baseras på att neuronätet har flera lager mellan input och output som gör det möjligt att lära sig den övergripande relationen mellan input och output i flera på varandra följande steg. Det gör att strategin som helhet blir mer korrekt och behöver mindre mänsklig vägledning.

Neuronät är bara ett maskininlärningsverktyg, men det finns många andra, med olika egenskaper: slumpskogar och förstärkta träd, klustermetoder, matrisfaktorisering osv.

En annan användbar typ av maskininlärningsstrategi kallas *förstärkningsinlärning*. Med denna strategi ger vi AI-systemet frihet att fatta beslut över tid och vid varje beslut ger vi systemet en belöningsignal som visar om beslutet var bra eller dåligt. Målet för systemet är att med tiden få så många positiva belöningar som möjligt. Denna strategi används t.ex. för rekommendationssystem (t.ex. den mängd webbaserade rekommendationssystem som ger användarna förslag om vad de kanske vill köpa) och marknadsföring.

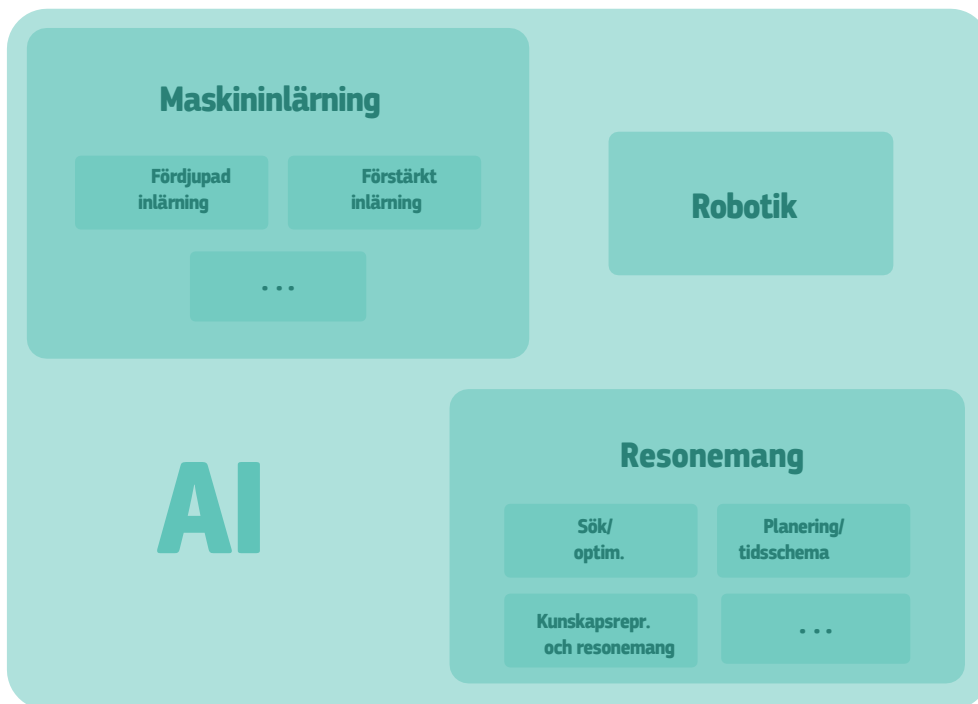
Maskininlärningsmetoder är inte bara användbara i perceptionsuppgifter som seende och textförståelse, utan i alla uppgifter som är svåra att definiera och inte går att beskriva uttömmande genom symboliska beteenderegler.

Lägg märke till skillnaden mellan maskininlärningsstrategier för att lära in en nu uppgift som inte går att beskriva väl på ett symboliskt sätt och rationella inlärningsagenter (som nämns i föregående avsnitt) som anpassar sitt beteende över tid för att uppnå målet bättre. Dessa båda tekniker kan överlappa eller samverka med varandra, men de är inte nödvändigtvis likadana.

Robotik. Robotik kan definieras som "AI i aktion i den fysiska världen" (kallas även *förkroppsligad AI*). En robot är en fysisk maskin som måste hantera dynamiken, osäkerheten och komplexiteten i den fysiska världen. Perceptions-, resonemangs, handlings- och inlärningsförmåga samt förmåga att interagera med andra system brukar vara integrerade i robotiksystemets kontrollarkitektur. Även andra discipliner än AI är viktiga för att konstruera och använda robotar, t.ex. maskinteknik och kontrollteori. Exempel på robotar kan vara robotmanipulatorer, självkörande fordon (t.ex. bilar, drönare, flygtaxi), humanoida robotar, robotdammsugare osv.

I figur 2 visas de flesta AI-underdiscipliner som anges ovan, samt hur de förhåller sig varandra. Tänk på att AI är mycket mer komplicerad än vad som framgår av den här bilden, eftersom den omfattar många andra underdiscipliner och tekniker. Som vi nämner ovan är robotiken också beroende av teknik som ligger utanför AI-området. Vi tror dock att detta är tillräckligt för att på ett fruktbart sätt kunna genomföra den process med att dela,

öka medvetenheten om och diskutera AI, AI-etik och AI-politik som måste äga rum i expertgruppen på hög nivå, som täcker in en mängd discipliner och berörda parter.



Figur 2: En förenklad översikt över underdisciplinerna inom AI samt deras inbördes förhållande.

Både maskininläring och resonemang omfattar många andra tekniker och robotik omfattar teknik som ligger utanför AI. AI som helhet ingår i disciplinen datorvetenskap.

3. Andra viktiga AI-begrepp och AI-frågor

Snäv (eller svag) och generell (eller stark) AI. Ett generellt AI-system är avsett att vara ett system som kan utföra de flesta aktiviteter som människor kan utföra. Snäva AI-system är i stället system som kan utföra en eller ett fåtal specifika uppgifter. De AI-system som används i dag är exempel på snäv AI. När AI först började utvecklas använde forskarna en annan terminologi (svag och stark AI). Det finns fortfarande många olösta etiska, vetenskapliga och tekniska problem med att bygga den kapacitet som skulle krävas för att uppnå generell AI, t.ex. sunt förnuft, resonemang, självmedvetande och maskinens förmåga att bestämma sitt eget syfte.

Datafrågor och snedvridning. Eftersom många AI-system, t.ex. de som omfattar övervakade maskininlärningskomponenter, är beroende av enorma mängder data för att fungera väl, är det viktigt att förstå hur data påverkar AI-systemets beteende. Om det t.ex. finns snedvridning i träningsdata, dvs. om datan inte är tillräckligt balanserad eller inkluderande, kommer ett AI-system som har tränats upp med den datan inte att kunna göra tillräckligt bra generaliseringar och kanske fattar orättvisa beslut som gynnar vissa grupper framför andra. På senare tid har AI-gemenskapen arbetat med metoder för att upptäcka och minska snedvridning i träningsdataset och även i andra delar av ett AI-system.

AI som en "svart låda" och förklarbarhet. Vissa maskininlärningsmetoder är visserligen mycket framgångsrika med tanke på korrekthet, men de är mycket ogenomskinliga när det gäller att förstå hur de fattar beslut. Begreppet *AI som en svart låda* avser sådana scenarier, där det inte går att följa spåret tillbaka till resonemanget bakom vissa beslut. Förklarbarhet är en egenskap hos AI-system som i stället kan ge en slags förklaring till sina åtgärder.

Målinriktad AI. Dagens AI-system är målinriktade, vilket innebär att en människa ger dem en specifikation på ett mål som ska uppnås och att de använder en viss teknik för att nå det målet. De fastställer inte sina egna mål. Vissa AI-

system (t.ex. de som baseras på vissa maskininlärningsmetoder) kan dock ha större frihet att avgöra vilken strategi som ska väljs för att uppnå det fastställda målet.

4. Uppdaterad definition av AI

Vi föreslår följande uppdaterade definition av AI:

”Artificiella intelligenssystem (AI) är programvarusystem (och eventuellt även hårdvarusystem) som har konstruerats av människor³ och som när de får ett komplext mål agerar i den fysiska eller digitala dimensionen genom att uppfatta sin omgivning via datainsamling och att tolka insamlade strukturerade eller ostrukturerade data, resonerar om den kunskap eller behandlar den information som härletts ur denna data och beslutar om den bästa åtgärd eller de bästa åtgärderna som ska vidtas för att uppnå det fastställda målet. AI-system kan använda symboliska regler eller lära sig en numerisk modell. De kan också anpassa sitt beteende genom att analysera hur miljön har påverkats av deras föregående åtgärder.

Som vetenskaplig disciplin innefattar AI flera metoder och tekniker, t.ex. maskininläring (som fördjupad inläring och förstärkt inläring är specifika exempel på), maskinresonemang (som omfattar planering, schemaläggning, kunskapsrepresentation och resonemang, sökning och optimering) och robotik (som omfattar kontroll, perception, sensorer och manöverdon samt integrering av alla övrig teknik i cyberfysiska system).”

Vi föreslår att detta dokument ska utgöra en referens och vara en källa till ytterligare information till stöd för denna definition.

³ Människor konstruerar AI-system direkt, men de kan också använda AI-tekniker för att optimera deras konstruktion.

Detta dokument har utarbetats av medlemmarna i expertgruppen på hög nivå för AI-frågor

som anges nedan i alfabetisk ordning

Pekka Ala-Pietilä, ordförande för AI HLEG AI Finland, Huhtamaki, Sanoma	Pierre Lucas Orgalim – Europas teknikindustri
Wilhelm Bauer Fraunhofer	Ieva Martinkenaite Telenor
Urs Bergmann Zalando	Thomas Metzinger JGU Mainz & European University Association
Mária Bielíková Slovak University of Technology in Bratislava	Cateljine Muller ALLAI Netherlands & EESC
Cecilia Bonefeld-Dahl DigitalEurope	Markus Noga SAP
Yann Bonnet ANSSI	Barry O’Sullivan, vice ordförande för AI HLEG University College Cork
Loubna Bouarfa OKRA	Ursula Pahl BEUC
Stéphan Brunessaux Airbus	Nicolas Petit University of Liège
Raja Chatila IEEE Initiative Ethics of Intelligent/Autonomous Systems & Sorbonne University	Christoph Peylo Bosch
Mark Coeckelbergh University of Vienna	Iris Plöger BDI
Virginia Dignum Umeå universitet	Stefano Quintarelli Garden Ventures
Luciano Floridi University of Oxford	Andrea Renda College of Europe Faculty & CEPS
Jean-Francois Gagné Element AI	Francesca Rossi* IBM
Chiara Giovannini ANEC	Cristina San José European Banking Federation
Joanna Goodey Europeiska unionens byrå för grundläggande rättigheter	George Sharkov Digital SME Alliance
Sami Haddadin Munich School of Robotics and MI	Philipp Slusallek German Research Centre for AI (DFKI)
Gry Hasselbalch The thinkdotank DataEthics & Copenhagen University	Françoise Soulié Fogelman AI Consultant
Fredrik Heintz Linköpings universitet	Saskia Steinacker Bayer
Fanny Hidvegi Access Now	Jaan Tallinn Ambient Sound Investment
Eric Hilgendorf University of Würzburg	Thierry Tingaud STMicroelectronics
Klaus Höckner Hilfsgemeinschaft der Blinden und Sehschwachen	Jakob Uszkoreit Google
Mari-Noëlle Jégo-Laveissière Orange	Aimee Van Wynsberghe TU Delft
Leo Kärkkäinen Nokia Bell Labs	Thiébaut Weber ETUC
Sabine Theresia Köszegi TU Wien	Cecile Wendling AXA
Robert Kroplewski Solicitor & Advisor to Polish Government	Karen Yeung The University of Birmingham
Elisabeth Ling RELX	

*Francesca Rossi har varit rapportör för detta dokument.