POLICYSAMMANFATTNING 2023:2

Vad menas med AI, vad regleras och varför är det viktigt?

Joakim Wernberg

Forskningsledare för Digitalisering och teknikpolitik vid Entreprenörskapsforum och lektor i Teknik och samhälle vid Lunds universitet

ENTREPRENÖRSKAPS FORUM

ENTREPRENÖRSKAPSFORUM

Entreprenörskapsforum är en oberoende stiftelse och den ledande nätverksorganisationen för att initiera och kommunicera policyrelevant forskning om entreprenörskap, innovationer och småföretag. Stiftelsens verksamhet finansieras med såväl offentliga medel som av privata forskningsstiftelser, näringslivs- och andra intresseorganisationer, företag och enskilda filantroper. Författarna svarar själva för problemformulering, val av analysmodell och slutsatser i rapporterna.

För mer information se www.entreprenorskapsforum.se

POLICYSAMMANFATTNING 2023:2

Vad menas med AI, vad regleras och varför är det viktigt?

Joakim Wernberg
forskningsledare för Digitalisering och teknikpolitik vid Entreprenörskapsforum
och lektor i Teknik och samhälle vid Lunds universitet
joakim.wernberg@entreprenorskapsforum.se

Tre aspekter av vad som menas med AI

Begreppet artificiell intelligens (AI) är i lika delar fantasieggande och förskräckande, inte minst för att det anspelar på maskiner som uppnår eller till och med överträffar mänsklig intelligens. Den bilden har nyligen förstärkts ytterligare av AI-modellen Chat-GPT, som över en natt har fått människor över hela världen att fastna i konversationer med en dator.

Dagens AI-tillämpningar skiljer sig emellertid fortfarande fundamentalt från både mänsklig och övermänsklig intelligens. Trots det är det långt ifrån alltid tydligt vad som faktiskt menas med AI i politiska diskussioner eller i samhällsdebatten. I skrivande stund håller ett förslag till AI-förordning (AI Act) på att slutförhandlas inom EU. Samtidigt pågår initiativ för att reglera AI på flera håll både inom och utanför Europa. Därför får definitionen av AI och AI-system en väldigt konkret innebörd som kommer att påverka innovation, entreprenörskap, digitala marknader och företagens regelbörda både på kort och lång sikt.

Med den reglering som nu står för dörren i EU ser det ut som att alltför många AI-tillämpningar, på grund av svårigheterna att definiera, klassas som högrisktillämpningar och därför kommer att omfattas av en överdimensionerad regelbörda. Utöver det riskerar regleringen att medföra tolkningssvårigheter och ökad osäkerhet på marknaden.

Det finns alltså goda skäl att titta närmare på vad som faktiskt menas med AI. I den här krönikan beskrivs tre olika aspekter av AI-begreppet som på olika vis bidrar till en bättre förståelse av vad AI kan och inte kan göra samt hur AI-system kan eller bör regleras.

1: Från deduktiv till induktiv ansats inom AI-forskningen

Ambitionen att bygga artificiell intelligens (AI) i form av en elektronisk, datorbaserad hjärna är långt ifrån ny utan kan spåras tillbaka till åtminstone 1950-talet.

Det etablerades två konkurrerande ansatser. Den första (deduktiva) ansatsen grundade sig på symbolhanterande maskiner som skulle kunna "tänka" deduktivt utifrån en representation av sin omvärld och resonera sig fram till en slutsats. Den andra (induktiva) ansatsen betonade induktivt "lärande" utifrån exempel inhämtade med någon typ av datainsamling i kombination med statistisk analys. Den deduktiva ansatsen hade initialt stort inflytande på AI-forskningen, men den AI vi pratar om idag bygger huvudsakligen på den induktiva ansatsen.

En starkt bidragande faktor till att den induktiva ansatsen har blivit dominerande är att den tekniska utvecklingen – datorkapaciteten och kanske framförallt tillgången till data – har kommit ikapp de teorier som ligger till grund för maskininlärning. Faktum är att många av de algoritmer som används i nya applikationer har funnits länge, men det är först nu det har varit möjligt att samla in och arbeta med tillräckligt stora mängder relevanta data för att tillämpa dem i praktiken (McAfee och Brynjolfsson, 2017; Wissner-Gross, 2017; Polson och Scott, 2018).

Polson och Scott (2018) pekar ut fyra steg som avgörande för genombrottet för den moderna maskininlärning som idag är mer eller mindre synonym med AI (Crawford, 2021). De fyra stegen består av:

- Komplexa modeller (sammansättningar av algoritmer):
  Utvecklingen har gått från små modeller som arbetar med enkla statistiska mönster till mer komplicerade sammansättningar av algoritmer för att beskriva komplexa mönster. Möjligheten att behandla en stor mängd parametrar parallellt har varit avgörande för denna utveckling.

- Stora och relevanta datamängder:
  För att få utväxling för komplexa modeller krävs tillräckligt stora och varierade datamängder. Annars riskerar modellen att överanpassa resultatet till befintlig datamängd, vilket försämrar prediktioner eftersom de bygger på mönster som är unika för just den datamängden.

- Experiment – att lära genom försök och misstag:
  Eftersom datormodellerna bygger på mönsterigenkänning i stora mängder data som ofta är oöverskådliga för människor finns det inte heller någon rättfram regel om hur de mönster som maskinen söker efter ska se ut. Modellen måste istället "lära sig" genom att successivt minimera felet i sina svar.

- Djupinlärning:
  Djupinlärning gör det möjligt att arbeta med och utvinna (mer) information ur komplexa indata genom att göra en stegvis analys. Modellen delar iterativt in data som ska klassificeras i en hierarki av kategorier (exempelvis pixel – öga – ansikte – person).

Det finns skäl att vara tydlig och konkret med vad som faktiskt avses när man pratar om AI inte minst därför att begreppet lockar till antropomorfiska tolkningar, det vill säga att maskiner tillskrivs mänskliga egenskaper som att ”tänka” eller ”lära sig”, som går långt utanför teknikens faktiska funktioner (Marcus och Davies, 2019).

Maskininlärning är en uppsättning metoder och verktyg för att utifrån någon form av indata identifiera och använda statistiska mönster för att göra förutsägelser om framtida utfall som i sin tur kan fungera som resultat eller indata för andra beslut och processer. En AI-modell som kan översätta engelska till svenska kan varken engelska eller svenska men har identifierat ett mönster för hur de två språken relaterar till varandra och använder det för att generera översättningar. Här skiljer sig maskinen väsentligt från en människa som lär sig ett språk i taget och därefter kan översätta mellan dem.

Den omskrivna AI-språkmodellen Chat-GPT som lanserades av OpenAI i slutet av 2022 har utöver att fungera som chattbot även använts till att skriva uppsatser, programmeringskod, låttexter och berättelser. Det betyder inte att Chat-GPT har bemästrat alla dessa områden och det finns gott om exempel på när modellen gör fel som framstår som triviala. Vad modellen kan är att identifiera strukturer inom en bred och nyanserad flora av olika kategorier av textbaserat innehåll.

En maskininlärningsmodells förmåga att leverera tillfredställande svar beror i hög grad på de data den utgår ifrån i sin induktiva analys. Det fungerar väl så länge den tillämpas i en kontext som inte förändras för mycket över tid och där etablerade statistiska mönster även framgent genererar önskvärda utfall. Om förutsättningar i modellens tillämpningsmiljö däremot förändras för mycket kan utfallet bli opålitligt och om historiska mönster inte är önskvärda att upprepa blir dess resultat otillfredsställande.

Med en mer konkretiserad beskrivning av AI blir det också tydligt att den nya tekniken inte är sprungen ur ett vacuum. Utöver de tillämpningar som ryms i AI-forskningens historia finns det också tydliga kopplingar till optimeringsteknik, reglerteknik och styrteori som alla används i industriella tillämpningar idag (Tillväxtanalys, 2022).

Chollet (2019) menar att AI-forskningen har konvergerat mot att estimera intelligens utifrån *uppvisad färdighet*, till exempel att spela brädspel eller köra bil, och *anpassningsförmåga* eller generaliseringsförmåga. Intelligensbegreppet i sig är svårdefinierat (Legg och Hutter, 2007) och båda dessa indikatorer beskriver intelligens indirekt utifrån förväntningar på hur en intelligent varelse – ofta används människan som jämförelse – skulle ha agerat eller presterat i en given situation. Dessa indikatorer är med andra ord mått på utfall och inte på intelligens. Samtidigt är det just intelligensbegreppet som förstärker både förhoppningar och oro i diskussioner om den artificiella intelligensens framtid. Därför är det viktigt att titta närmare på hur olika typer av AI förhåller sig till intelligens.

2: Smal AI, generell AI och superintelligens

De tillämpningar av AI som finns idag beskrivs ofta som smal AI. De kan utföra en väldigt väl avgränsad och smal typ av uppgift enormt väl, ofta långt bättre än människor. Men utanför sin tillämpningsdomän förlorar de sin förmåga och blir snabbt ineffektiva, opålitliga eller obrukbara (Domingos, 2015; Mitchell, 2019).

En typ av AI-modeller som har kommit att utmärka sig är s.k. *foundation models* eller *general-purpose AI models*. Det är stora AI-modeller i termer av både beräkningskapacitet och insamlad träningsdata som fungerar som en centraliserad infrastruktur för en avgränsad kategori av AI-tillämpningar, särskilt inom språkbehandling (Bommasani m.fl., 2021). För AI- modeller där marginalnyttan av data inte avtar eller avtar långsamt finns det betydande stordriftsfördelar i att skapa en central modell som andra aktörer sedan tar del av som en tjänst, anpassar för egna tillämpningar eller använder som input i egna modeller. Framväxten av dessa stora modeller ger också en stark indikation om hur det håller på att växa fram nya värdekedjor baserade på data och informationsbehandling i ekonomin. Dessa värdekedjor kommer att vara centrala för att realisera nyttan med datadriven innovation.

Ett av de främsta uttalade målen för dagens AI-forskning är generell AI (Artificial General Intelligence eller AGI) som till skillnad från smala AI-modeller förmår att anpassa sig inte bara till förändringar inom ett tillämpningsområde utan också till nya tillämpningsområden. Det råder oenighet inom forskningen huruvida tillräckligt stora sammansatta modeller med flera smala AI-tillämpningar kan bli en AGI-modell, eller om det finns en mer fundamental kvalitativ skillnad mellan smal AI och AGI (se exempelvis Brockman, 2020). Dessa meningsskiljaktigheter letar sig tillbaka till skilda perspektiv på vad intelligens är och hur vi mäter det (Chollet, 2019).

Bortsett från den mer filosofiska diskussionen om relationen mellan smal AI och generell AI, är det inte otänkbart att det kommer att bli svårt att avgöra när en AI-modell faktiskt uppvisar generell artificiell intelligens. Beroende på vilken typ av definition av AGI man väljer måste en AGI-modell antingen kunna göra allting som en människa kan eller vara tillräckligt generaliserbar över en avgränsad mängd tillämpningsområden. Är till exempel de stora språkmodeller (foundation models) som finns idag att betrakta som AGI? Oaktat vilken måttstock man väljer, kommer den att visa sig svår att implementera på ett entydigt vis eftersom det som ska påvisas är intelligens och det som mäts är utfallet av intelligens. Även om en modell klarar av de uppsatta målen kommer frågan vara om den uppvisar generell intelligens eller bara är begränsat generaliserbar. Därför är det sannolikt att vi även med avsevärda framsteg i utvecklingen av AGI kommer att befinna oss i en definitionsmässig gråzon.

Bortom åtskillnaden mellan smal och generell AI finns det också de som föreställer sig framväxten av superintelligens som uppnår och överträffar mänsklig intelligens i alla

avseenden (Kurzweil, 2014; Bostrom, 2014). Det som skiljer den här typen av artificiell intelligens från smal eller generell AI är inte bara dess generaliserbarhet i tillämpningar utan också att den tillskrivs någon form av medvetande eller självmedvetande som utgångspunkt för hur den agerar. De flesta inom AI-relaterad forskning tycks vara överens om att dagens smala AI-tillämpningar inte på egen hand eller med inkrementell utveckling kan bli självmedvetna superintelligenser.

Det pågår i skrivande stund både forskning och debatt om hur man bäst kan förebygga de potentiellt existentiella risker som förknippas med superintelligensincidenter (Bostrom, 2014; Tegmark, 2018; Hawking, 2018; Russell, 2019). Ur detta perspektiv präglas debatten om AI av försiktighetsprincipen och kan närmast liknas vid hanteringen av kärnvapen eller annan teknik vars riskhantering utgår från potentiellt omfattande negativa konsekvenser.

Skillnaderna mellan smal AI och AGI är påtagliga även om de med tiden kan visa sig svåra att precisera och skillnaderna mellan dem och superintelligens är i sin tur fundamentala. Trots det tenderar de att blandas ihop i den samhällspolitiska debatten om AI.

Alla tre former av AI – smal, generell och superintelligent – har som mål att efterlikna olika aspekter av intelligens, men det är inte osannolikt att de med tiden kommer att bygga på delvis eller helt olika typ av underliggande teknik. Precis som AI-begreppets innebörd har förändrats sedan 1950-talet och gammal teknik har ersatts med ny kan vi sannolikt räkna med att det händer igen. Vad betyder det för hur vi definierar praktiska tillämpningar av AI?

3: AI-definitioner i politik och samhällsdebatt: intelligens, teknik eller arbete?

AI-utvecklingen har sedan 1950-talet visat på en begreppsförändring som innebär att gammal teknik har lyfts ut och ny teknik har lyfts in i vad som avses med AI (McCorduck, 2004; Crawford, 2021). Det är en naturlig utveckling inom forskningen, men kan visa sig betydligt mer problematiskt när begreppet lyfts ut och används i en bredare samhällsdebatt eller i politik och lagstiftning (Larsson, 2020). Det tar sig uttryck i påtagliga svårigheter att definiera och avgränsa vad ett AI-system, det vill säga praktiska tillämpningar av AI, är i till exempel insamling av statistik, politiska främjandeinitiativ eller utformningen av ny reglering.[1]

Statistiska Centralbyrån har översatt en tidig definition från EU-kommissionen i syfte att samla i statistik om AI-användning som lyder (SCB, 2020):

1. I formuleringen av nya regelverk är man noga med att inte definiera AI i generell bemärkelse utan specifikt vad som avses med ett AI-system.

> *Artificiell intelligens (AI) syftar till system som uppvisar intelligent beteende genom att analysera sin omgivning och agera, med någon nivå av självbestämmande, för att uppnå specifika mål. AI-baserade system kan vara ren mjukvara eller inbyggda i hårdvara.*

EU-kommissionen (2021) introducerade en annan definition i förslaget till en ny AI-förordning:[2]

> *"Artificial intelligence system" (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with.*

Den första definitionen betonar "intelligent beteende" och "agerande" men är teknikneutral. Den andra definitionens betydelse styrs i hög utsträckning av en lista av specifika mjukvarubaserade tekniker som den hänvisar till. OECD (2019) har formulerat en tredje definition som istället betonar det arbete som ett AI-system utför:

> *An AI system is a machine-based system that is capable of influencing the environment by producing an output (predictions, recommendations or decisions) for a given set of objectives. It uses machine and/or human-based data and inputs to (i) perceive real and/or virtual environments; (ii) abstract these perceptions into models through analysis in an automated manner (e.g., with machine learning), or manually; and (iii) use model inference to formulate options for outcomes. AI systems are designed to operate with varying levels of autonomy.*

EU-kommissionens oberoende expertgrupp inom AI tog fram en mer detaljerad definition, som kommissionen frångick i förslaget till AI-förordning, som liknar OECD:s definition och betonar den artificiella intelligensens funktion och det arbete den utför på ett så teknikneutralt sätt som möjligt (AI HLEG, 2019):

> *Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems*

2. Definitionen har under arbetet med AI-förordningen varit föremål för mycket debatt och är i skrivande stund fortfarande inte entydigt bestämd, vilket vittnar om dess betydelse.

> *can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.*

Vid första anblick kan de olika definitionerna ovan tyckas snarlika, men de betonar tre olika innebörder av AI-system med utgångspunkt i intelligens, teknik respektive arbete.

Det är olyckligt när definitioner av AI – formella såväl som informella – utanför forskningen förlitar sig på intelligensbegreppet eftersom det är svårt att precisera, avgränsa och mäta. Det inbjuder till antropomorfiska tolkningar och lämnar ett ofta överdrivet tolkningsutrymme kring hur maskinens agens ska betraktas i förhållande till människans.

Samtidigt som det kan vara eftersträvansvärt att formulera teknikneutrala regelverk riskerar de att bli så generella att de är svåra att tolka och därför får en bredare tillämpning än avsett. Regelverk som avgränsas utifrån en tydlig uppsättning tekniska tillämpningar riskerar istället att skapa en obalans i regelbörda mellan de tillämpningar som omfattas av reglerna och de som faller utanför dess avgränsning.

Detta talar sammantaget för att arbeta med definitioner som frångår intelligensbegreppet och istället betonar maskinens funktion eller det (kognitiva) arbete den utför. En sådan definition skapar goda förutsättningar för att diskutera ansvar, ägande och reglering av AI och AI-baserade tjänster. Samtidigt blir det enklare att identifiera och avgränsa fall där kognitivt eller analytiskt arbete behöver regleras på ett särskilt sätt om det utförs av en maskin.

Varför spelar det roll vad vi menar med AI?

Vilken definition som används för att främja, reglera eller mäta AI-tillämpningen i ekonomi och samhälle har stor betydelse för vilken typ av teknik och verksamheter som kommer att omfattas i praktiken. Det finns en gränsdragningsproblematik som riskerar att skapa incitament för att *inte* få ett system kategoriserat som AI eller att ”fördumma” AI-baserade tjänster innan de introduceras på den europeiska marknaden för att undvika extra regelbörda. I värsta fall ökar incitamenten för att förlägga innovation och entreprenörskap på AI-området utanför EU.

Ännu mer problematiskt blir det förstås om olika överlappande regelverk utgår från olika definitioner eller om olika grupper i regleringsarbetet utgår från uppfattningar som grundar sig i olika definitioner. Det riskerar att medföra onödigt stor regelbörda, tolkningssvårigheter och ökad osäkerhet på marknaden. Ur ett marknadsperspektiv är det viktigt att etablera tydliga och långsiktiga spelregler för marknaden att förhålla sig till samt se till att den resulterande regelbördan är proportionerlig till regleringens syfte. Det är inte självklart att de nya AI-regleringar som nu arbetas fram lever upp till dessa krav.

Vidare behandlas inte skillnaden mellan smal AI, AGI och superintelligens explicit i någon av de olika definitionerna av AI-system.[3] Det kan tas till intäkt för att ytterligare utvecklingssteg inom AI-området ska kunna inrymmas politiska regelverk som nu tas fram idag, men det riskerar också att medföra betydande problematik. Det är utmanande att formulera regelverk som ska fungera både för den teknik vi har idag och den teknik vi kan tänkas få i framtiden. Till exempel kan det skapa oförutsedda effekter och onödig regelbörda för de som arbetar med tillämpningar av smal AI om de framgent kommer att omfattas av regelverk utformade för AGI eller till och med superintelligens. En alternativ ansats vore att reglera dessa olika typer av AI vart och ett för sig. Det skulle till exempel underlätta arbetet med att identifiera och avgränsa högrisktillämpningar av AI.

Med den reglering som nu står för dörren i EU ser det ut som att många AI-tillämpningar som ligger på spektrumet mellan smal AI och AGI att klassas som högrisktillämpningar och omfattas av ytterligare regler som i allt väsentligt snarare är dimensionerade för superintelligens. Dessutom finns det en risk att de som bygger applikationer ovanpå stora AI-modeller (foundation models eller general-purpose AI models) som Chat-GPT kommer att betraktas som utvecklare av nya högrisktillämpningar och omfattas av motsvarande reglering. Reglering som är dimensionerad för kärnvapen tillämpas då på teknik med en helt annan typ av riskprofil.

Regleringen av AI, tillsammans med tidigare och kommande regelverk för dataskydd och dataflöden, kommer att spela en allt viktigare roll för det europeiska näringslivets konkurrenskraft i takt med att ekonomin blir allt mer tjänsteintensiv och tjänsterna blir allt mer mjukvarubaserade och datadrivna. I skrivande stund behandlas frågor om AI och data allt för ofta som vore det avgränsade eller branschspecifika fenomen, men inom överskådlig framtid kommer datadrivna värdekedjor genomsyra hela ekonomin. Det är jag inte övertygad om att de nya regelverken håller för.

Referenser

AI HLEG (2019). A definition of AI: Main capabilities and disciplines. Idependent High-level Expert Group on Artificial Intelligence, set up by the European Commission.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. a*rXiv preprint arXiv*:2108.07258.

3. Den nämns dock i EU-kommissionens oberoende expertgrupps underlag till den definition de formulerade

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies Reprint Edition. Oxford University Press.

Brockman, J. (Ed.). (2020). *Possible minds: Twenty-five ways of looking at AI*. Penguin.

Chollet, F. (2019). On the measure of intelligence. *arXiv preprint arXiv*:1911.01547.

Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books.

EU-kommissionen (2021). Proposal for a regulation of the European parliament and of the council: Laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. Com(2021) 206 Final.

Hawking, S. (2018). *Brief answers to the big questions*. Bantam.

Kurzweil, R. (2014). *The singularity is near* (pp. 393-406). Palgrave Macmillan UK.

Larsson, S. (2020). On the governance of artificial intelligence through ethics guidelines. *Asian Journal of Law and Society*, 7(3), 437-451.

Legg, S., & Hutter, M. (2007). A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and applications*, 157, 17.

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Vintage.

McAfee, A., & Brynjolfsson, E. (2017). *Machine, platform, crowd: Harnessing our digital future*. WW Norton & Company.

McCorduck, P. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. CRC Press.

Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans*. Penguin UK.

OECD (2019), "Scoping the OECD AI principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO)", *OECD Digital Economy Papers*, No. 291, OECD Publishing, Paris, https://doi.org/10.1787/d62f618a-en.

Polson, N., & Scott, J. (2018). *AIQ: How artificial intelligence works and how we can harness its power for a better world*. Random House.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.

SCB (2020). Artificiell intelligens i Sverige. Statistiska Centralbyrån.

Tegmark, M. (2018). *Life 3.0: Being human in the age of artificial intelligence*. Vintage.

Tillväxtanalys (2022). Varför AI? – Förutsättningar, möjligheter och hinder för företag att använda AI. Tillväxtanalys rapport 2022:11.

Wissner-Gross, A. (2017), "Data sets over Algorithms" i Know this – Today's most interesting and important scientific ideas, discoveries and developments, Brockman, J. (ed), HarperCollins Publishers, New York.


ENTREPRENÖRSKAPS
FORUM