Artificiell intelligens & mänsklig hjärna

Artificiell intelligens (AI) i olika former finns redan här i vår vardag, men har på senare tid fått ett starkt fokus i nyhetsflöde och samhällsdebatt. I denna artikel av **Magnus Johnsson**, universitetslektor och docent i datavetenskap vid Malmö universitet, får vi en spännande insyn om hur artificiell intelligens fungerar jämfört med den mänskliga hjärnan.

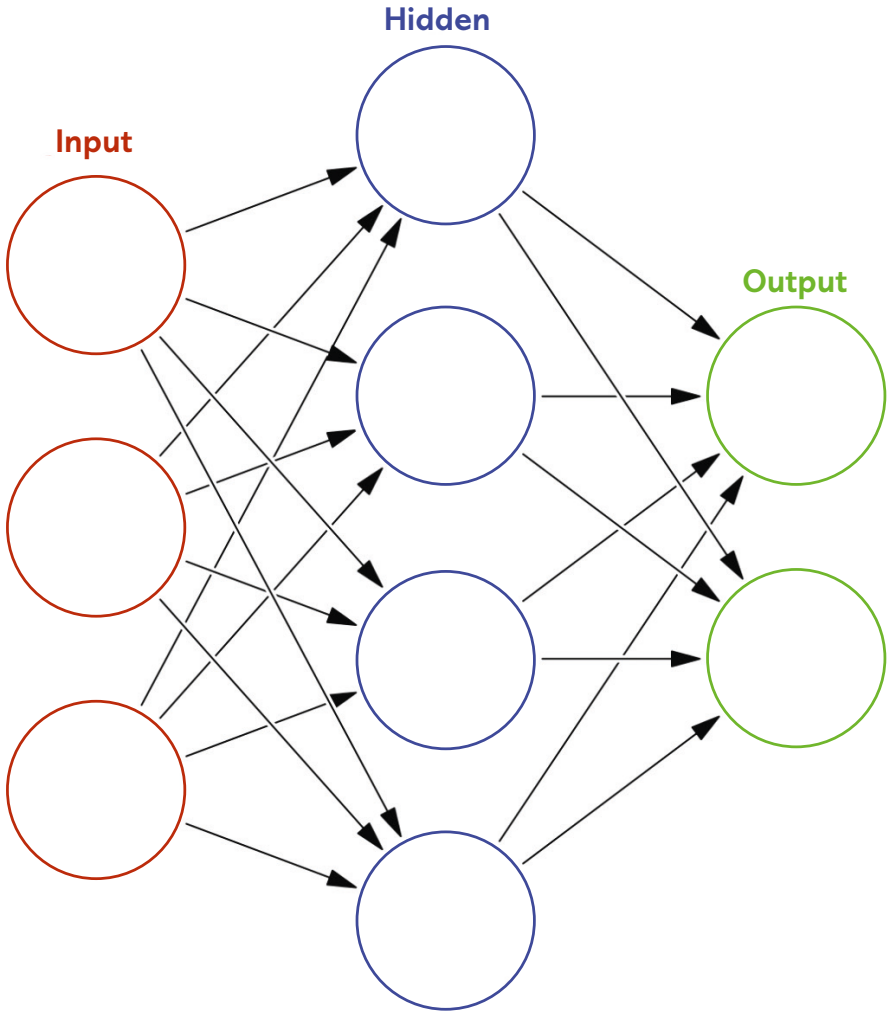
Människan har i alla tider försökt förstå hur det kan komma sig att vi kan tänka, känna, uppleva och manipulera vår omvärld. Detta har efterhand resulterat i en mängd olika perspektiv, skolor och områden, till exempel filosofi, kognitionsvetenskap, psykologi och neurovetenskap som mer systematiskt, men med lite olika infallsvinklar, försöker svara på dessa frågor.

Artificiell intelligens – AI – går bortom detta och försöker skapa mekanismer som helt eller delvis reproducerar vissa av dessa förmågor. Precis vilka beror lite på vem man frågar och hur området avgränsas. Vad som menas med AI är därför inte helt enkelt att definiera. Detta beror delvis på att begreppet ”intelligens” som sådant inte är enkelt att ge en klar och av alla accepterad definition. För AI kompliceras det ytterligare genom att olika forskare och praktiker har olika uppfattningar om vad som är bra utgångspunkter och tillvägagångssätt för att åstadkomma AI. Ett vanligt synsätt är dock kortfattat att inom AI studeras hur man kan skapa maskiner – i praktiken vanligtvis datorprogram – som är kapabla till sådant som skulle kräva intelligens om en människa skulle göra det.

ETT TVÄRDISCIPLINÄRT OMRÅDE

AI är ett starkt tvärdisciplinärt område och idéer, perspektiv och tekniker har lånats in från många områden, bland annat filosofi, matematik, ekonomi, neurovetenskap, psykologi, datavetenskap och lingvistik. En följd av detta är att ett antal olika sätt att närma sig problemet med att implementera AI har utkristalliserats. Standardförfarandet när man utvecklar ett program som kan spela schack eller andra brädspel baseras till exempel på spelteori (matematik och ekonomi) där programmet söker efter det bästa draget under antagandet att motståndaren också spelar


766.71
614.00



Ett artificiellt neuronnät är en sammankopplad grupp noder, likt det stora nätverk av neuroner i en hjärna. Varje nod i bilden representerar en artificiell neuron och en pil representerar en koppling från utdatan av en neuron till indatan av en annan neuron.

optimalt, medan de bästa programmen för bildigenkänning i dag baseras på AI-metoder som hämtat sin inspiration från neurovetenskapen.

Vi har på senare tid fått se ett starkt fokus på AI och maskininlärning, som är ett delområde inom AI, i nyhetsflödet och den samtida debatten. Som så många gånger tidigare lyfts en ny teknologi fram som både frälsare och förgörare. Egentligen är AI dock inte något nytt område, utan har växt fram under de senast cirka sjuttio åren. Det som satt fokus på det under senare tid är vissa ”genombrott” när det gäller till exempel bildigenkänning och chatbottar.

Faktum är att dessa ”genombrott” inte är en konsekvens av i första hand några nya principiella insikter, utan mer beror på att man nu på grund av betydligt mer data och kraftfullare hårdvara har kunnat skala upp tillämpningen av vissa AI-metoder som har varit kända i flera årtionden. Det är i första hand artificiella neuronnät det handlar om.

Artificiella neuronnät är en hel klass av algoritmer som är mer eller mindre inspirerade av naturliga neuronnät. Jag ska här beskriva hur ett enklare sådant kan fungera i princip, samtidigt som jag jämför med ett naturligt neuronnät. Liksom naturliga neuronnät består ett artificiellt neuronnät av ”neuron” (det vill säga matematiska modeller) som är sammankopplade med varandra. Insignalerna till ett neuron i ett naturligt neuron motsvaras i ett artificiellt neuronnät av indata i form av reella tal. Synapserna i ett naturligt neuronnät motsvaras i sin tur av parametrar, det vill säga ytterligare reella tal, kallade vikter. Varje neuron kan ha flera indata och dessa summeras efter att de har multiplicerats med sina respektive vikter. Om den viktade summan överstiger ett tröskelvärde så ger neuronet en utsignal (=1), annars 0 som kan ses som motsvarigheten till ingen utsignal (av skäl som har att göra med implementationen av inlärningen i artificiella neuronnät approximeras denna ”trappfunktion” i praktiken med en snarlik matematisk funktion som är differentierbar, till exempel en så kallad sigmoidfunktion). Detta motsvarar, i en abstrakt mening, uppbyggandet och propageringen av en aktionspotential i neuronen i ett naturligt neuronnät. I ett naturligt neuronnät är modifikation av synapsernas egenskaper en avgörande komponent vid inlärning. I ett artificiellt neuronnät motsvaras detta av en modifikation av vikterna. Modifikationen av vikterna bestäms av en inlärningsalgoritm som försöker ändra vikterna i neuronnätet givet en uppsättning träningsdata (som består av exem-

Detta motsvarar, i en abstrakt mening, uppbyggandet och propageringen av en aktionspotential i neuronen i ett naturligt neuronnät.

> *Dessutom finns det ett värde genom att vi lär oss om naturlig kognition inte enbart genom ett analytiskt tillvägagångssätt, utan även genom syntes, det vill säga genom att försöka skapa den artificiellt.*

pel på indata och motsvarande korrekt utdata). Vikterna i neuronnätet modifieras av algoritmen på ett sådant sätt att felet mellan den faktiskt producerade utdatan och den korrekta utdatan för hela träningsmängden minimeras.

I moderna artificiella neuronnät är neuronen ofta arrangerade i många lager, där varje lager, givet riktningen på informationsflödet, utvecklar representationer på en högre ”abstraktionsnivå” Motsvarigheter till detta finner vi i hjärnan, till exempel ventrala och dorsala synströmmarna. En oro som finns hos en del debattörer är att AI snart ska gå bortom och bli överlägsen mänsklig intelligens. Faktum är att detta redan skett för länge sedan, inom snäva domänspecifika områden. Exempelvis räknar en kalkylator snabbare än människor, och för de allra flesta av oss kan upplevelsen av att spela schack mot en schackdator jämföras med känslan av att försöka springa ikapp med en motorcykel. Man kan skilja på AI som är mer eller mindre domänspecifik eller mer generell AI.

DOMÄNSPECIFIK AI OCH GENERELL AI

Domänspecifik AI löser ett snävt AI-problem. Ett program som kan spela schack, men inget annat, är ett exempel på detta. Generell AI innebär en mer generaliserad förmåga att lösa problem, som kräver intelligens inom många domäner, såsom en människa kan. Förutom domänspecifik AI och generell AI kan man föreställa sig allt däremellan, och hypotetiskt bortom generell AI på en mänsklig nivå, det vill säga super-generell-AI. Faran nu sägs vara att vi har genombrott inom AI som är mindre domänspecifik, alltså mer generell. Detta beror förmodligen till stor del på de chatbottar som dykt upp den senaste tiden med en till synes imponerande och bredare förmåga. Denna förmåga beror dock inte i första hand på några principiella genombrott, utan på en uppskalning av algoritmer (såsom artificiella neuronnät) som i princip varit kända i årtionden, där man använder maskininlärningsalgoritmer som tränats på stora delar av den data som finns på internet.

Det är skalan som bländar, men de problem som finns hos dessa chatbottar – till exempel att man inte kan lita på om svaren är korrekta eller om chatbotten ”hallucinerar” ett felaktigt svar som är tillsynes trovärdigt för den som inte har expertkunskap på området, till exempel för att man ställt frågor som extrapolerar utanför träningsmängden – kommer inte att kunna lösas enbart med mer av samma metoder.

Frågan är då hur man går vidare för att skapa en generell AI, om detta är önskvärt och om det ens går. Detta är nämligen inte självklart. Till exempel har Landgrebe och Smith (2023) gett en mycket omfattande argumentation för att även om generell intelligens hos människan helt och hållet är en konsekvens av hjärnans materiella struktur och funktion, så är det av matematiska skäl inte möjligt att modellera den på en sådan nivå som möjliggör generell AI.

En annan aspekt på problemet som inte är helt förstådd är det mänskliga medvetandet. Om det är avgörande för en generell AI att ha ett subjektivt medvetande (qualia) är oklart. Och eftersom problemet med medvetandet inte tillfullo är förstått, är det av det skälet oklart om det ens i princip går att implementera en generell AI. Penrose (1989) argumenterar till exempel för att mänskligt medvetande är icke-algoritmiskt och om det är korrekt, så går det inte att implementera på en dator. Det finns åtminstone anledning att inte helt utesluta att om vi som människor har ett subjektivt medvetande (det finns argument för att vi inte har det, fastän vi tror det), så är det inte orimligt att tänka sig att det har en funktion som ger oss en bättre överlevnadsförmåga, och att det kanske är avgörande för generell intelligens.

Huruvida jag tror på dessa argument mot generell AI eller inte låter jag vara osagt. Oavsett så finns det ett värde i att forska om generell AI, eftersom AI-tillämpningarna som erhålls på vägen och som redan uppnåtts är landvinningar som har stort värde och som har stor potential till positiva förändringar för våra liv och för världen. Dessutom finns det ett värde genom att vi lär oss om naturlig kognition inte enbart genom ett analytiskt tillvägagångssätt, utan även genom syntes, det vill säga genom att försöka skapa den artificiellt. Särskilt om man använder sig av ett biologiskt inspirerat tillvägagångssätt.

För att gå vidare i utvecklingen mot generell AI eller åtminstone mer generella AI-system så är mitt tillvägagångssätt att, liksom när det gäller artificiella neuronnät, inspireras av biologin och naturen för att koppla ihop ett flertal artificiella neuronnät till mer sammansatta arkitekturer. Detta tillvägagångssätt bygger på ett antal biologiskt inspirerade principer som jag tror är väsentliga för att efterlikna olika förmågor i däggdjurshjärnan, såsom perception, förväntningar, inre föreställningar, minne, förmåga att representera nya begrepp och att föreställa sig fiktiva och även omöjliga föremål. Jag utvecklar detta i Johnsson (2022).

MAGNUS JOHNSSON
Universitetslektor och docent i datavetenskap, fil dr i kognitionsvetenskap, Malmö universitet
magnus.johnsson@mau.se & magnus@magnusjohnsson.se

Referenser

Johnsson M (2022). Perceptions, Imagery, Memory and Consciousness, Polish Journal of Science and Philosophy. Philosophical and Interdisciplinary Studies. ISSN 2545-1936.

Landgrebe J & Smith B (2023). Why Machines Will Never Rule the World, Routledge, ISBN 978-1-032-31516-4.

Penrose R (1989). The Emperor's New mind: Concerning Computers, Minds and The Laws of Physics, Oxford University Press, ISBN 0-19-851973-7.