VARFÖR FINNS DET INGA RIKTIGA ROBOTAR?

Peter Gärdenfors och Christian Balkenius

Kognitionsforskning, Lunds Universitet,
S–223 50 Lund
e-mail: peter.gardenfors@fil.lu.se
christian.balkenius@fil.lu.se

Här är först en högtidlig programförklaring: Kognitionsforskning kan beskrivas som studiet av hur information representeras och bearbetas i naturliga system, i synnerhet den mänskliga hjärnan, och hur detta kan modelleras i datorer och andra artificiella system. De centrala forskningsområdena för kognitionsforskningen är perceptions- och minnesmodeller, kunskapsrepresentation, inlärning, begreppsbildning, språkförståelse, problemlösning, planering, samt interaktion mellan människa och dator. Ett övergripande mål är att förstå de kognitiva processernas funktion och hur de kodas i hjärnan, men forskningen strävar också efter att kunna simulera dessa processer med hjälp av datorer.

Varför finns det då inga robotar av den sort man träffar på inom science fiction, dvs. robotar som självständigt och smart kan utföra komplicerade uppgifter? Om kognitionsforskningen kan uppnå de allmänna mål som givits ovan, så kommer vi att förstå hur människan löser problem och hur detta kan simuleras med datorer och då bör det vara möjligt att bygga en riktig robot. Men dit har vi tydligen inte nått ännu. Problemet att bygga en robot är ett bra testproblem för kognitionsforskningen eftersom en sådan behöver perception, minne, kunskap, inlärning, kommunikations- och planerings-

förmåga, dvs. just de kognitiva färdigheter som forskningen sysslar med enligt programförklaringen. De s.k. robotar man träffar på inom industrin har mycket litet av detta – de kan utföra en snäv repertoar av rutinuppgifter i en strikt anpassad miljö. Vill man ändra deras beteende måste de vanligtvis programmeras om. De är med andra ord, osjälvständiga, oflexibla och har anpassningssvårigheter.

Naturen har, genom evolutionens tålmodighet, lyckats lösa dessa problem med eleganta metoder. De flesta djur utvecklas till självständiga individer, ofta med en underbar anpassningsförmåga. Ställda inför nya typer av problem i helt främmande omgivningar, lyckas djuren normalt finna beteenden som är lämpliga, eller åtminstone inte direkt skadliga. Varför skall det vara så svårt att konstruera en robot som imiterar naturens lösningar? Varför kan vi inte bygga en maskin som är lika kapabel som en kackerlacka?

Man finner naturliga självständigt handlande varelser, eller *autonoma agenter* som de kallas inom artificiell intelligens (AI), i djurvärldens hela hierarki, från de encelliga toffeldjuren till människan (om nu människan står högst). De enkla djurens beteende klassificeras inom AI som *reaktiva system*. Detta innebär att de inte har någon framförhållning utan reagerar på stimuli allt eftersom de dyker upp. De uppvisar vad som brukar kallas ett stimulus-respons-beteende.

En grov beskrivning av arkitekturen för ett reaktivt system, som passar in på en behavioristisk syn på agenter, består av flera komponenter eller ”moduler”. Först och främst behövs en *perceptuell modul* som hanterar de stimuli agenten får genom olika receptorer. Hos djur är detta de olika sinnesorganen; hos en robot kan det vara en uppsättning sensorer eller en videobild. För att agenten skall kunna göra något krävs en *beteendemodul* som styr agentens motorik, dvs. hur den rör sig, och eventuella andra former av ”responser” (t.ex. om den ger ifrån sig ljud- eller doftsignaler). Hos djur sker detta vanligen genom signaler till olika muskelgrupper; hos en robot kan det

ske genom styrning av olika motorer eller ljudgeneratorer. Kopplingen mellan den perceptuella modulen och beteendemodulen styrs av en *reaktiv modul*. Denna modul svarar mot den ”svarta låda” som behaviorister brukar tala om.

En faktor som oftast glöms bort inom AI är att djur och människor har en *motivation* för sitt beteende. Från evolutionsteorins synvinkel är de yttersta målen överlevnad och reproduktion, men dessa mål är vanligtvis inte medvetna eller ens representerade hos individerna. Motivationen är ett system som väljer beteende i olika situationer på ett sätt som i stort sett överensstämmer med evolutionens krav. Eftersom individen har flera behov och därmed konkurrerande mål, behöver den ett smidigt system som väljer mellan olika handlingar så att sökandet efter ett mål inte förstör möjligheten att uppfylla ett annat. Det skall också vara möjligt för agenten att tillfälligt bortse från ett mål för att uppfylla ett annat.

En reaktiv agent behöver därför också en *motivationsmodul* som för varje tillfälle styr in det på ett speciellt mål. Modulens roll är att bedöma agentens aktuella behov, mål och möjligheter så att den kan välja vilken typ av beteende som för tillfället har störst chans att uppfylla något av agentens mål. Motivationssystemet är också ansvarigt för bedömningen av agentens beteende så att det kan förändras på olika sätt beroende på om det är framgångsrikt eller inte. En ytterligare roll för motivationssystemet är att jämföra förväntade belöningar med de verkliga belöningarna efter olika beteenden och att förändra dessa förväntningar när de inte stämmer med den aktuella omgivningen.[1]

En ensam agent måste kunna fatta alla beslut själv. I enkla världar där ingenting förändras kan sådana beslut vara fast programmerade i agenten. I en föränderlig omgivning måste en autonom agent hela tiden anpassa sig till de nya förhållandena.

---

[1]En modell av ett sådant motivationssystem presenteras i Christian Balkenius’ uppsatser ”The roots of motivation” som publiceras i *From animals to animats II*, Cambridge, MA: MIT Press, 1993 och ”On motivation” (manuskript).

Förändringar som kan uppstå består t.ex. i att vägarna till de olika målen förändras eller blockeras. Det är också möjligt att mål förändras t.ex. genom att de byter värde eller förflyttar sig. Det är viktigt att agenten kan anpassa sig och leta upp nya sätt att nå målen eller att hitta alternativa mål. En sådan agent blir mycket mer komplicerad än sin förprogrammerade motsvarighet eftersom den måste uppvisa *adaptivt beteende*.

För att hantera denna förmåga krävs att systemet kompletteras med en eller flera *adaptiva moduler* som kontrollerar den ursprungliga reaktiva agenten. En viktig poäng är att denna modul ”läggs ovanpå” det reaktiva systemet, som finns kvar, i stort sett oförändrat, för att ta hand om ”rutinartat” beteende. Det adaptiva systemet förutsätter således ett reaktivt system för att kunna fungera.

I ett reaktivt system sker inlärningen enbart genom betingning. Men den adaptiva modulen gör det möjligt för agenten att lära sig nya beteenden på ett mer avancerat sätt. På ett liknande sätt består minnesfunktionen i ett reaktivt system bara av att värdena på olika parametrar kan förändras, medan ett adaptivt system kan hantera andra typer av minnen. En adaptiv agent kan *göra* saker med minnena. Ett reaktivt system kan lära sig att hitta i en labyrint, medan det krävs ett adaptivt system för att kunna ta genvägar om några av labyrintens väggar tas bort. Härmed får vi ett system som går utöver de behavioristiska, eftersom denna typ av minne inte kan förklaras med deras modeller.

Nästa steg mot en fullvärdig autonom robot består i att förse en adaptiv agent med en förmåga att *planera*. För att kunna göra detta måste agenten kunna förutse konsekvenserna av sina handlingar. Det kräver i sin tur att den har någon form av *modell* av sin omgivning. I den enklaste formen kan denna ses som enbart något som ger feedback till händelser som representeras i systemet. Om händelserna representerar möjliga handlingar agenten kan utföra, så ger modellen en uppfattning om vilka konsekvenserna skulle bli om handlingen

utfördes i den ”yttre” miljön. En agent med en inre modell av världen kan, med andra ord, *simulera* konsekvenserna av olika beteenden.[2] Som ett sista steg i arkitekturen läggs därför en *modellmodul* ovanpå det adaptiva systemet. På samma sätt som ovan förutsätter denna modul ett fungerande adaptivt system.

Inom traditionell AI har man gärna utvecklat system som enbart är avsedda att planera men som inte har de mer fundamentala förmågor som ett reaktivt eller adaptivt system har. Till skillnad från denna strategi anser vi att planering utgör det sista steget i konstruktionen av en autonom agent och inte det första och att detta steg i hög grad är beroende av de andra.

Poängen med en modellmodul är att man kan låta ett trial-and-error- beteende fortgå i denna i stället för i den yttre obönhörliga verkligheten. För att en modell skall kunna fungera förutsätts något fundamentalt nytt, nämligen *representationer* av information. Det räcker inte med att informationen om exempelvis en fara kan nå agenten och ge upphov till ett undflyende beteende, utan agenten måste kunna utnyttja informationen, *föreställa sig* faran, även när den inte är fysiskt närvarande för att den inre modellen skall kunna vara till någon nytta. Här spelar minnet en avgörande roll: Från det kan systemet konstruera lämpliga representationer till den inre modellen. *Hur* representationerna skall se ut är en av de mest fundamentala frågorna för kognitionsforskningen.

Möjligheten att föreställa sig en handling och dess konsekvenser, i stället för att direkt utföra den, är en nödvändig betingelse för att man skall kunna planera.[3] Genom att representera olika handlingar, olika vägar som kan leda till målet, så får agenten *valmöjligheter*. Utan representationer av flera alternativ kan man inte välja.

Avancerad planering innefattar inte bara planering av enskilda handlingar utan också *sekvenser* av handlingar. Först representeras en handling i modellmodulen, varefter konsekvenserna simuleras. Och i den nya modellsituation som uppstår på detta sätt representerar man

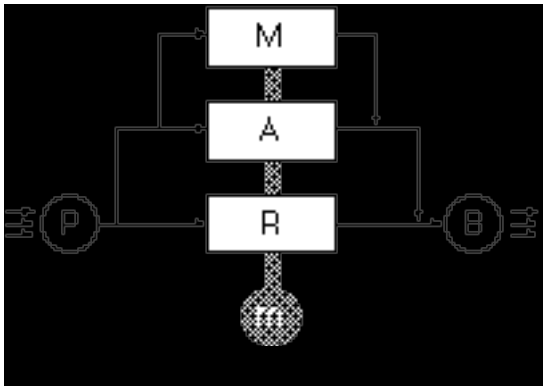
---

[2]För en ytterligare diskussion av den inre modellens roll, se Peter Gärdenfors: ”Medvetandets evolution”, Kapitel 5 i *Blotta tanken*, Nya Doxa 1992.

[3]En utförlig analys av de evolutionära och systemteoretiska grunderna för planering ges i Agneta Gulz’ avhandling *The Planning of Action as a Cognitive and Biological Phenomenon*, Lund University Cognitive Studies; 2, 1991.

en efterföljande handling och simulerar i sin tur nya konsekvenser, etc.

Vi har därmed nått fram till en skiss av funktionerna hos en autonom agent, som ser ut på följande sätt:



Arkitekturen hos en autonom agent: (P) Perceptuellt system, (R) Reaktiv modul; (B) Beteendemodul; (A) Adaptiv modul; (M) Modellmodul; (m) Motivationsmodul.

Denna arkitektur kan jämföras med hjärnans uppbyggnad. De evolutionärt äldre delarna av hjärnan styr kroppens grundläggande funktioner. Exempelvis fungerar delar av det limbiska systemet som en motivationsmodul. Perceptionsmodulen svarar mot delar av hjärnbarken, t.ex. visuella cortex. Motorik och andra former av beteende styrs av lillhjärnan. Modellmodulen svarar närmast mot den evolutionärt färska främre delen av hjärnbarken där tidsmedvetande och planering är lokaliserade. Den föreslagna arkitekturen för en autonom agent har således en viss biologisk rimlighet.

Varför kan då inte dagens kognitionsforskning konstruera en robot som har de egenskaper som ges av den arkitektur vi föreslår? För att förstå varför vi ännu inte kommit dit måste vi återvända till kognitionsforskningens grundvalar. Det finns, grovt talat, två konkurrerande synsätt på de fundamentala kognitiva processerna och deras arkitektur. En skola ser kognition som *symbolhantering*. Man antar att all information i hjärnan kan representeras med någon symbolisk kod och att denna behandlas enligt ett system av regler. Detta synsätt på tänkandet har varit dominerande inom filosofin under större delen av seklet och de flesta modeller inom lingvistik och AI använder sig av en sådan kod. Om man uppfattar hjärnan som ett

informationsbehandlande system, så ligger det nära till hands att jämföra den med en dator. De flesta framgångsrika program inom AI bygger på en symbolisk representation av informationen. Ny kunskap uppnås genom att tillämpa logiska slutledningsregler på den givna informationen.

Den dominerande traditionen bland AI-forskare har varit att se hjärnan som en likartad ”symbolhanterare”. Förespråkarna för detta synsätt hävdar att tänkande, problemlösande, språkförståelse och andra kognitiva funktioner bygger på slutledningar som utförs med hjälp av satser i ett mentalt språk. Hjärnan beskrivs alltså som en traditionell dator där man tänker sig att programmet som gör slutledningarna finns lagrat i hjärnsubstansen.

Men ju mer man lär sig om hur kognition fungerar hos människor och djur, desto tydligare blir det att den symboliska representationen inte är en lämplig form. Om vi ser på hur hjärnan representerar information, så kommer den symboliska formen evolutionärt sist. AI har i en mening börjat med att bygga skorstenen till ett hus som saknar grund. Man har koncentrerat sig på de funktioner som finns i modellmodulen, snarare än att utgå från beteendet och hur det hanteras i reaktiva och adaptiva system. AI och kognitiv psykologi har också byggt på de olika delarna som om de vore oberoende av varandra. Man har inom kognitiv psykologi studerat minne, inlärning och perception som oberoende funktioner och mycket litet intresserat sig för deras interaktion.

Det kan tyckas att den arkitektur vi har föreslagit för en autonom agent också är uppbyggd av ett antal oberoende moduler. Men ingen av modulerna kan fungera självständigt utan enbart i *interaktion* med andra komponenter. Det finns exempelvis ingen modul som är ansvarig för planering utan denna förmåga uppstår endast genom samverkan mellan de olika modulerna.

Vidare fungererar inte det mänskliga minnet som i en traditionell dator där minnesenheter lagras i avgränsade och oberoende celler. Minnen i biologiska system är inte oberoende enheter som de blir med den symboliska representation som används i en dator. Ett minne är alltid knutet till den kontext där minnet uppstod. Dessutom får vi olika typer av minnen beroende på vilken modul vi betraktar. Perceptuella minnen, exempelvis, är beroende av den perceptuella modulens funktion. Minnena uppstår genom att parametrarna i modulen

ackommoderas efter perceptionerna. Minnena i beteendemodulen, å andra sidan, lagras som associationer mellan en typ av situation och en handling. Man antar att associationerna uppstår genom klassisk betingning.

Inom AI bestäms systemets beteende framför allt med hjälp av en modell av omvärlden som ges en symbolisk representation. Inlärning beskrivs som förändringar av modellen. Representationen i modellen är emellertid oberoende av perception och beteende. Eftersom man inte vet hur perceptuella och motoriska koder skall kopplas till de symboliska blir det svårt att få systemet att lära sig genom perceptioner och att översätta vad det lärt sig till handling. En AI-modell har inga behov och därför kan den inte heller ha någon motivation. I stället konstrueras ett AI-system att uppfylla ett mål som bara *beskrivs* i den symboliska notationen.

I en autonom agent av den typ vi förespråkar är inlärning direkt kopplad till perception och handling. De representationer som uppkommer vid inlärningen är alltså inte oberoende av agentens beteenden. Detta gör att dess representationer av omvärlden är anpassade till de handlingar som den faktiskt kan utföra. Inlärning är således adaption av beteende istället för förändring av en omvärlds-modell.

Den omvärldsmodell vi arbetar med är inte en beskrivning av världen som i AI utan *en värld i sig*. Agenten kan lära sig genom att göra saker i denna inre värld med samma mekanismer som när den lär sig genom handlingar i den yttre världen. Det är denna form av inlärning som ligger bakom problemlösning, planering och fantasi.

Eftersom dessa och liknande problem framstår som alltmer hopplösa för den symbolorienterade traditionen har det utvecklats en alternativ skola (med rötter i associationistisk psykologi) som arbetar med de informationskoder som finns i perceptuella och motoriska system. För att bygga datoranpassade modeller av processerna i sådana utnyttjar man framför allt s.k. *artificiella neuronnät*.

Det finns flera skäl varför neuronnät är mer lämpade att modellera de olika funktionerna hos en autonom agent. Först och främst arbetar de direkt med perceptionerna i den form som matas in. Den perceptuella informationen behöver alltså inte kodas om. De ger också rika möjligheter att modellera inlärning. Beroende på vilken typ av neuronnät som används kan man fånga olika typer av biologisk

inlärning, t.ex. associativ inlärning, och kategorisering. Ett symbolbaserat system kraschar i många situationer, medan ett neuronnät alltid ger någon form av output. Om detta output är olämpligt, kommer den att lära sig detta och reagera bättre nästa gång. En annan fördel jämfört med de klassiska systemen är att när väl nätverket lärt sig en uppgift utförs den mycket snabbt utan omfattande beräkningar. När en autonom agent tränats för en uppgift krävs bara det reaktiva systemet för att utföra den. Även små skador på ett symbolbaserat AI-system leder till total kollaps av systemet, medan ett neuronnät normalt fortsätter att fungera även om en del neuroner slås ut. Nätverket kan till och med anpassa sig till sina skador och därmed kringgå dem.

Av dessa skäl tror vi att de artificiella neuronnäten utgör de rätta byggbitarna för en autonom agent med den arkitektur som skisserats ovan. Men vi vill betona att på grund av modulernas varierande funktioner kommer det att bli nödvändigt att använda flera olika typer av neuronnät och olika inlärningsmekanismer. Om man inom kognitionsforskningen lyckas konstruera neuronnät för de funktionella modulerna vi diskuterat, så finns det gott hopp om att kunna göra riktiga robotar.