Peter Ekberg

Referens inom mänsklig kognition och AI

Den brittiske AI-pionjären och logikern Alan Turing (1912–1954) föreslog i sin berömda text "Computing machinery and intelligence" att hjärnan var en digital beräkningsmaskin och att en människa vid födseln är en oorganiserad "maskin" som via träning organiseras till en högre "universell" struktur kapabel att lösa de mest intrikata problem. Eftersom mänskligt tänkande ansågs ske beräkningsmässigt ställde Turing upp ett test som han kallade "the imitation game" för att försöka svara på om även en digital maskin, en dator, kan vara förmögen att tänka. Turings imitationstest går ut på att någon (x) får föra ett samtal med datorn[1] och en verklig person (y) han inte är bekant med. Om x inte kan avgöra vem som är datorn och vem som är y då klarar datorn testet och anses kunna tänka.[2] Nedan är ett utdrag ur Turings text där han presenterar en fiktiv diskussion mellan utfrågaren Q och datorn A. Lägg märke till att datorn är programmerad att efterlikna mänskligt beteende. Det finns mängder med verkliga människor som aldrig för sitt liv skulle kunna tänka sig att skriva en dikt och motsvarande mängd som säkerligen skulle räkna fel på räkneexemplet.

Q. Var vänlig och skriv en sonett på temat the Forth Bridge.

A . Räkna inte med mig. Jag har aldrig kunnat skriva poesi.

Q. Addera 34957 till 70764.

A. (avvaktar i ungeär 30 sekunder och ger sedan svaret) 105621.[3]

[1] Liknande maskiner finns det gott om idag, även om ingen ännu kan sägas ha klarat Turingtestet. De kallas "chatterbots" och jag vill passa på att rekommendera en trevlig chatterbot vid namn Alan som du kan träffa på www.a-i.com/ .

[2] Detta är givetvis omdiskuterat, Searle, Putnam m.fl. menar att tänkande kräver förståelse för vad man gör, inte bara förmåga att rätt kunna sätta samman, för en själv, betydelselösa symboler.

[3] A. Turing, "Computing machinery and intelligence", 1950. Datorn är här klurig nog att ge fel svar (ett typiskt mänskligt drag), det korrekta svaret är 105721.

Filosofisk tidskrift 2004 nr 1, 49-53

Jag skall i denna artikel diskutera den amerikanske filosofen och Harvardprofessorn Hilary Putnams (1926–) version av Turingtestet som i boken *Reason, truth and history* (1981)[4], omformuleras till att bli ett test för referens, vilket han då kallar Turings referenstest. Testerna liknar varandra, men i referenstestet är inte längre målet att avgöra om den vi samtalar med är en verklig person eller en maskin, utan att avgöra om den vi konverserar med använder orden för att *referera* som vi gör.

Ponera att vi genomför testet med en "chatterbot" som klarar referenstestet. Den får klart godkänt tack vare att den ledigt och lätt (via text på dataskärm) konverserar på ett naturligt språk och tycks referera till den vanliga typen av föremål, kan svara på kluriga ontologiska och vardagspraktiska frågor. Borde vi då inte kunna dra slutsatsen att maskinen refererar till föremål på samma sätt som vi? Turing skulle förmodligen svara ja på den frågan, men Putnam säger bestämt nej, det kan vi inte! Det må vara ett bra test, men det är inte för den skull nödvändigt att vi har gemensam referens. En maskin kan klara referenstestet utan att referera till någonting alls! Och Putnam anser att det är precis vad de symbolbehandlande formella systemen inom AI gör.[5] Chatterboten uppfyller inga av kraven som ställs upp för referens. Den saknar inte bara elektroniska ögon och öron, den är inte förberedd eller mottaglig för att bearbeta ett inflöde från dessa organ och den vet inget om kontroll av en kropp. Visst kan den samtala med oss på flytande engelska om det underbara landskapet i New England, men den skulle aldrig kunna känna igen någon av ingredienserna i beskrivningen, kor, ängar, kyrkor etc, även om den var placerad på stora ängen framför kyrkan en solig söndag. Datorprogrammets felfria syntaktiska satser och utsagor står, hur anmärkningsvärda de än må vara, inte på något sätt i förbund med den verkliga världen! Vi kan alltså inte lägga större vikt vid datorns prat om New England än myrans slumpmässiga streck i sanden som en avbild av Churchill. Ingendera *refererar* till någonting. Så långt allt väl.

[4] Putnams text behandlar huvudsakligen en applikation av referensteorin på en variant av Descartes klassiska kunskapsteoretiska problem och vill därmed visa att en viss form av skepticism (världen är en illusion) är begreppsligt omöjlig.

[5] Filosofen John Searle kommer till liknande slutsats i sin klassiska text "Minds. Brains and programmes" (1980). Det är viktigt att påpeka att både Putnam och Searles uppsatser är från tidigt åttiotal, där AI-projektet gick ut på fullständig förprogrammering av agentens egenskaper och kunskaper. Föreliggande uppsats behandlar delvis de framsteg som gjorts inom området och som föranleder oss rätta våra teorier och vår kritik efter det nya paradigmet. Filosofer är viktigare än någonsin för AI projektet!

1. SPRÅKLIGA INGÅNGS- OCH UTGÅNGSREGLER

Det är ju vi människor som betraktare och samtalspartners till datorn som upprätthåller en representationskonvention, vi tolkar de ord chatterbot-datorn producerar *som om* de refererade till verkligheten, vilket skapar en illusion av att datorn lyckas referera och därmed också mena någonting med vad den säger och uppträda allmänt intelligent. Putnam menar att vi människor kan referera till objekt när vi talar, skriver eller tänker genom vår kognitiva direktkontakt med tingen vi möter i världen. En maskin kan det *inte* eftersom maskinens representation av världen inte grundar sig i någon som helst perceptuell erfarenhet. Här kommer vi till den centrala poängen i Putnams diskussion om referens inom AI kontra mänsklig referens. När människor talar om tingen i sin omgivning så är detta tal nära förknippat med våra icke-verbala relationer (perceptioner, bilder, minnen etc.) till tingen i omvärlden. Det finns språkliga *ingångsregler* som leder oss från en erfarenhet av någonting i verkligheten till ett yttrande om tinget. Det finns också omvänt språkliga *utgångsregler* som leder oss från beslut uttryckta i språk och ord till icke-språkliga handlingar (först formulerar vi vad vi skall göra och sedan gör vi det). Eftersom ett AI-system i Putnams världsbild saknar både språkliga ingångs- och utgångsregler så utför den ingenting annat än ett syntaktiskt arbete baserat på den kunskapsrepresentation vi gett systemet med hjälp av logik och algoritmer. Ur vår synvinkel *liknar* det arbete chatterboten utför mänsklig intelligens och referens.

2. DET NYA AI-PARADIGMET

Under de sista fem-tio åren har det skett en metodologisk förskjutning inom AI-projektet. Det finns i världen idag system, rationella agenter, som inte enbart behandlar förprogrammerad information, utan har en växande erfarenhetsbas som de tillskansar sig genom interaktion med omgivningen. Dessa agenter kan sägas ha både syn, hörsel och hjärna. De är system som lär sig att orientera sig i världen och utvecklar sin förståelse av omvärlden genom erfarenhet som inhämtas från artificiella sinnen.

I boken *Artificial Intelligence – A modern approach* (Norvig och Russell 2003) definieras AI på följande sätt: "AI är studiet av agenter som tar emot percept från omgivningen och utför aktioner."

"Tar emot percept från omgivningen" är nyckelorden här och det är även i dessa ord det stora metodologiska brottet ligger. AI-systemen är inte längre nödvändigtvis icke-referande enheter avskurna från verkligheten, förprogrammerade och evigt dömda att blint behandla meningslösa symboler enligt formella regler. De stöts och blöts mot omgivningen

enligt "trial and error" principen. Ett fönster mot omvärlden har alltså öppnats enligt det nya paradigmet och det är min uppfattning att vissa AI-system idag kan sägas referera naturligt till objekten de möter i sin omgivning![6] Putnam skulle säkert hålla med om detta. Han ställer upp två krav för referens: (1) Agenten måste ha en *avsikt* med sin referens, en myra som råkar avbilda Winston Churchill när den släpar sig fram i blöt sand refererar inte till Churchill. (2) Agenten måste också stå i *direktkontakt* med världen. Hjärnor som i näringslösningsbad frikopplade från kroppen stimuleras till upplevelser av ondsinta vetenskapsmän refererar inte till någonting även om det för hjärnan verkar som att den är en person som visslande knallar i vårsolen över universitetets grönområde.

Dessa båda kriterier är uppfyllda av de AI-system vi diskuterar här. I fotnot 6 nämndes "The ALVINN computer vision system". Överfört i Putnams terminologi kan vi säga att ALVINNs *avsikt* är att förhålla sig till och processa ett inflöde av vägpercept för att kunna framföra fordonet utan skada. ALVINN står också i *direktkontakt* med yttervärlden.

På liknande sätt kan de nya systemen sägas ha språkliga ingångs- och utgångsregler realiserade och de utnyttjar sin perception, erfarenhet, minne etc. för att förhålla sig autonomt till världen därefter på ett meningsfullt sätt. Jag tror att den principiella skillnaden mellan AI-systemens "kognitiva" processer och informationsbehandling och våra egna drastiskt har minskat tack vare det nya paradigmet, även om mänsklig kognition fortfarande är vida överlägsen även de allra bästa lärande systemen. Det är på sin plats att vi blir varse att den här delen av Putnam och Searles kritik mot AI inte längre är aktuell. Deras argument är mycket bra och riktiga för sin tid. Men de diskuterade GOFAI (Haugeland, 1985), där man trodde sig kunna förprogrammera agentens hela kunskapsbas och beteende. De allra flesta AI-forskare är idag ense om att GOFAI är en väg som bara leder in i återvändsgränder. Världen är helt enkelt för komplex. Det finns alltför många faktorer att ta hänsyn till för att vi skall kunna förprogrammera all information agenten behöver. Och vad händer om det inte finns några givet *rätta* beslut att fatta när man ändå *måste* fatta ett beslut? AI-världen har gått vidare och vi bör sålunda bredda vår kunskap om projektet. För nu är filosoferna viktigare än någonsin om AI i stark mening[7] skall kunna realiseras. En annan viktig poäng är den mängd arbete som finns att göra *innan* autonoma agenter blir en del av

[6] Exempel är " The ALVINN computer vision system", som autonomt navigerade en minivan 2 850 miles genom USA. Eller inom kirurgin där robotassistenter som skapar bildrepresentationer av omgivningen (och aldrig darrar på handen), börjat användas.

[7] Verkligt intelligenta, medvetna artificiella agenter.

samhällets vardag. Exempelvis frågor rörande etiska och moraliska implikationer för artificiella agenter. Hur skall de implementeras i samhället? Skall de ha mänskliga rättigheter och skyldigheter? Hur skall agenternas världsbild se ut? Dessa frågor behöver svar och vi vill självklart vara med och påverka utvecklingen. Vi kan börja arbetet med att uppdatera oss om AI-projektet och tillerkänna agenterna förmåga att kunna referera till objekten de möter i världen även om deras perceptuella processer är realiserade i ett annat medium än våra egna.

LITTERATUR

John Haugeland. 1985. *Artifical Intelligence – The very idea.* MIT press paperback edition 2000.

Thomas Mautner. 2000. *The penguin dictionary of philosophy.* Penguin books.

Peter Nordin. 2003. *AI, Artificiell intelligens och intelligenta robotar* (preliminär version av bok under utgivning hösten 2003 studentlitteratur).

Peter Norvig och Stuart Russell. 2003. *Artificial Intelligence – A modern approach.* Second edition.

Hilary Putnam. 1981. *Reason truth and history*, Inledningskapitlet till denna bok, "Hjärnor i näringslösning", finns i *Filosofin genom tiderna*, efter 1950, 2:a rev. uppl., utg. av Konrad Marc-Wogau, Lars Bergström, Staffan Carlshamre. Thales, 2000.

Searle J. R. 1980. "Minds, brains and programs". *Behavioral and Brain Sciences*, 3, 417–457.

Turing, Alan. 1950. "Computing Machinery and Intelligence". *Mind*, 59, s. 434–460.