ROBOTAR OCH LIV

Det finns en oro att robotar ska bli intelligentare än människor, och rent av medvetna. Men exakt vad är det vi (bör) vara oroliga för? **Lars Hertzberg** föreslår att genom att skilja på teknisk och organisk intelligens kan en viktig skillnad rörande önskningar och behov tydliggöras. Liknande sammanbladningar förekommer kring frågan om vem (eller vad) som är medveten.

Text: Lars Hertzberg

Intelligens

I januari i år publicerades ett upprop (”*Research Priorities for Robust and Beneficial Artificial Intelligence: an Open Letter*”, finns publicerat på futureoflife.org) undertecknat bland annat av fysikern **Stephen Hawking**, den svensk-brittiske forskaren **Nick Boström** och ett antal experter på artificiell intelligens, där man varnade för följderna av en okontrollerad utveckling av intelligenta robotar och andra datorsystem. Robotarna kunde bli ett hot mot mänsklighetens fortsatta existens, resonerade man. Problematiken summeras på följande sätt av **Björn Sundell** (HBL 2015-06-01):

Intelligensen har varit människans konkurrensfördel i evolutionen. Den har gett homo sapiens makt över naturen och den har banat väg för alltmer sofistikerad teknik. Vad händer när den tekniken blir allt intelligentare, ja, intelligentare än människan själv?

Här utmålas människan som evolutionens krona, som ”vinnare” i ”kampen för tillvaron”. Sundell verkar ha satt fingret på det tänkande som ligger bakom forskarnas manifest. I och för sig är varningar för teknologier som utvecklas bortom all kontroll befogade. Men den form forskarnas bekymmer tar sig verkar ogenomtänkt. Det rör sig inte om en vetenskaplig föreställning. Ur biologins synvinkel är framgång alltid något relativt: en art är bättre anpassad än en annan till en viss livsmiljö. Att tala om en arts framgång i absolut mening är en begreppslig förvirring. (Möjligen kunde man betrakta en arts långvariga fortbestånd som ett allmänt mått på framgång – i så fall är kackerlackorna förmodligen en av de mest lyckade arterna.)

Man kunde för all del använda ordet ”intelligens” som en sammanfattande beteckning på många av de överlevnadsstrategier som är specifika just för arten människa: utvecklade samarbetsformer, förmågan att planmässigt forma om miljön (bygga bostäder, odla grödor osv), att utnyttja den (redskap, jaktvapen) och att uthärda den (kläder, uppvärmning, läkekonst, konservering av mat). Talorganen och händerna har här varit centrala tillgångar. ”Intelligensen” i denna mening har bland annat gjort det möjligt för människor att – till skillnad från många andra arter – bosätta sig över stora delar av jordklotet, och att specialisera sig på varierande livnäringsformer. Man kan tala om en differentiering *inom* arten.

Men att utmåla människans relation till robotarna som en tävlan mellan intelligenser verkar vara uttryck för en tankemässig förvirring. Människans intelligens är en aspekt av hennes sätt att leva. Den kommer bland annat till uttryck i hennes förmåga att komma till rätta med de praktiska problem tillvaron ställer henne inför. Kanske kunde man tala om en ”organisk intelligens”.

Robotarnas intelligens är däremot inte en del av ett liv. Problemlösande robotars hela tillvaro *består*, kunde man säga, *i att lösa problem* och inget annat. Robotarna själva har inte några problem, deras lösningar är inte till för dem. De problem robotar sätts att lösa definieras av de människor som programmerar dem. Visserligen kan robotar utvecklas till att programmera om sig själva – de ska, i en viss mening, kunna lära sig av erfarenheten. Men detta gör dem inte till problemägande varelser. Vi kunde säga att de har en ”teknisk intelligens”.

”Människans intelligens är en aspekt av hennes sätt att leva.”

Eric Horvitz, som är forskningsdirektör vid Microsoft, skriver: ”en dag kan vi förlora kontrollen över system av artificiell intelligens genom uppkomsten av superintelligenser som inte handlar i enlighet med mänskliga önskemål” (citerat i det ovan nämnda upropet). Horvitz verkar ha blandat ihop två olika farhågor. Att vi kan förlora kontrollen över datorsystem är självklart. Det vet var och en som i vardagen använder mobiltelefoner eller datorer. Också experterna kan förlora kontrollen. Från senaste millennieskifte minns vi att många experter sa sig inte veta vad som skulle hända när datorernas klockor passerade årtusendegränsen. Visserligen inträffade inga större mankemang, men osäkerheten var av-

slöjande. Man visste inte om datorerna skulle fortsätta att fungera som man önskade sig. Kanske det vore riktigare att säga: vi kan inte vara säkra på att datorer alltid fungerar som vi *hoppas*. Här finns en obegränsad potential för kaos.

Den andra farhågan som skymtar fram i Horvitz formulering är däremot att datorerna ska börja fungera i enlighet med *andra* önskningar än de mänskliga. Han och de andra forskarna verkar tycka att det är den verkligt skrämmande tanken. Frågan är då varifrån dessa andra önskningar skulle härstamma. Tanken tycks vara: om datorerna inte fungerar enligt våra önskningar så måste de fungera enligt sina egna – men det är inte en hållbar slutledning.

Tankefelet uppstår genom att man uppfattar robotarna som något slags levande varelser. Man stirrar sig blind på det man föreställer sig att människor och robotar har gemensamt: intelligensen, och bortser från den radikala skillnaden mellan organisk och teknisk intelligens. Här rör det sig om en begreppslig klyfta som inga teknologiska framsteg kan överbrygga (lika lite som problemet med tidsresor kan övervinnas genom framsteg, säg, inom bilindustrin).

Förmodligen är man rädd för att robotarna kan komma att drivas av något slags självbevarelsedrift: de strävar efter att ”överleva” och kanske också att mångfaldiga sig själva. På så sätt skulle vi ställas inför en ny variant av den darwinistiska kampen för tillvaron. (Kanske tänker man sig kampen för tillvaron som en fortsättning på schackmatchen mellan superdatorn **Deep Blue** och världsmästaren **Garry Kasparov**.) Men där brister analogin: människors och djurs problemlösningsförmåga växer fram ur överlevnadens behov – hos robotarna tänker man sig att deras överlevnadsbehov på något sätt skulle genereras ur deras problemlösningsförmåga.

Den radikala skillnaden mellan organisk och teknisk intelligens [är] en begreppslig klyfta som inga teknologiska framsteg kan överbrygga

Detta tema har berörts i fiktionen. I **Stanley Kubrick**s film *2001: A Space Odyssey* (1968) är det datorn **HAL 9000** som försöker ta herraväldet över en rymdfarkost för att undgå att stängas ner. I **Alex Garland**s *Ex machina* (2015) är det ett datorsystem, inkarnerat (om det är det rätta ordet) som en skön kvinna i plast och metall vid namn **Ava**, som förför en man för att lura honom att rädda ”henne” från att ersättas med ett mera avancerat datorsystem. Båda filmerna innehåller mycket som är tänkvärt (jag har skrivit om *Ex machina* på min blogg ”Language is the thing we do”), men det gäller att minnas att det rör sig om fiktion. Fråga dig till exempel *vad* föremålet för datorsystemets överlevnadsinstinkt i Ex machina skulle kunna vara. Är det inkarnationen Ava – alltså den specifika ansiktsmask och peruk, de specifika kläder och metallskelett i vilka datorns mekanismer höljts in? Eller är det den enorma databas som lagrats i datorns hårdskiva? Eller är det själva mjukvaran? Och i så fall: exakt vari består mjukvarans överlevnad? Innebär övergången till Windows 10 att Windows 8.1 i någon mening har gått i graven? Snarare är det en vidareutveckling ur det tidigare programmet, liksom mitt 10-åriga jag inte innebar att mitt 8-åriga jag hade förintats. Men i Hawking & Co:s scenarier försöker dataprogrammen inte kämpa för sin överlevnad, tvärtom bygger rädslan på att de kommer att ersätta sig själva med mera sofistikerade program.

Kort sagt: i föreställningen om en kamp om tillvaron mellan människor och robotar saknas det en uppfattning vems eller vilkets tillvaro robotarna antas försvara. Vad vi har att frukta är inte robotarnas intelligens, utan deras eventuella brist på intelligens, det vill säga konstruktörernas begränsade förståelse för sina skapelser.

Medvetande

Debatten om maskiners förmåga att tänka fick en viktig impuls av **Alan Turing**s artikel ”Computing Machinery and Intelligence” (*Mind* 1950, finns tillgänglig på nätet). Där lanserade han det så kallade *Turingtestet* eller *imitationsspelet*: försökspersonerna får skriva ner frågor och läser svaren. De vet inte om svaren kommer från en dator eller en människa. Om datorn i fem minuter lyckas få 30 % av försökspersonerna att tro att den är en människa har den klarat testet. Det betyder att den – enligt den definition Turing föreslår – kan tänka.

Man kan fundera på poängen med det här testet. Turings definition är givetvis en godtycklig stipulation. Ingenting hindrar oss att introducera nya sätt att använda ett ord som ”tänka” – illusionen är att tro att man därigenom sagt något nytt om vad tänkande *är*. Huruvida vi vill använda ordet ”tänka” i fråga om datorer är närmast en praktisk fråga. (I min ungdom skilde man mellan tänkande och vanliga hissar: tänkande hissar kunde hålla ordning på flera knapptryck och stanna vid de olika våningarna i tur och ordning. Där hade ordet ”tänkande” en funktion.)

Ibland ställs frågan: kan datorer utvecklas därhän att de har ett medvetande? Turing ställer frågan men hans behandling av den är valhänt (artikeln i övrigt är helt frejdig och läsbar om än inte speciellt djupgående). Han säger likt många efter honom att om man vill vara säker på om en maskin har medvetande måste man själv vara maskinen och känna efter – det solipsistiska argumentet. Han tror att de flesta av oss inte väljer den linjen, utan är villiga att godta imitationsspelet som ett kriterium på tänkande.

Hela frågan kring maskiners medvetande – det sätt att använda ordet ”medvetande” som aktualiseras här – tycks mig förvirrad. Vad är det att ”ha medvetande”? Har människor medvetande? Vi säger om levande varelser att de befinner sig vid varierande grader av medvetande respektive medvetslöshet (människor och många djur kan vara vakna, dåsa till, somna in, vara avsvimmade, ligga i koma). Att vara vid medvetande är ett tillstånd hos organismen, inte en artegenskap. När det gäller datorer verkar vi inte ha behov av att tala om sådana här tillstånd. (Min dator ”fryser till” med

jämna mellanrum, men jag är inte benägen att säga att den ”förlorar medvetandet” – snarare att den ”blir förlamad”.)

I debatten har medvetandet däremot upplyfts till en mystisk egenskap som olika typer av varelser eller anläggningar likmätigt sin natur antas ha eller sakna, en egenskap vi har en direkt upplevelse av hos oss själva men inte har tillgång till hos andra. Den föreställningen bottnar, tror jag, i en begreppslig förvirring.

Jag föreställer mig att en – och kanske den viktigaste - av de saker man försöker uttrycka när man talar om medvetande bättre kunde uttryckas genom att tala om att vara ett subjekt: det vill säga att vara någon som gör bedömningar och utför handlingar. En bedömning – till exempel ”Den här tomaten går inte att äta, den är för mjuk” – eller en handling – till exempel att slänga bort tomaten – är något jag kan göras ansvarig för. Mitt beslut kan diskuteras, någon kan kritisera

I debatten har medvetandet upplyfts till en mystisk egenskap som olika typer av varelser eller anläggningar likmätigt sin natur antas ha eller sakna

det, fråga efter mina skäl, jag kan försvara mig eller gå med på att jag misstog mig osv. Som subjekt rör jag mig i en sfär av ställningstaganden: klokt eller dumt, rätt eller fel. Att vara subjekt – att vara ansvarig – är inte något jag bara är bekant med från mitt eget fall, tvärtom: det är människorna i min omgivning som lär mig förstå vad det innebär att ha ansvar. När jag växer upp lär jag mig själv delta i formandet av uppfattningarna om hur saker och ting ska bedömas, hur man kan handla eller ska handla i olika situationer. Jag är ett subjekt genom att tillhöra ett samfund av subjekt.

Om detta är vad vi menar med frågan efter medvetandet, blir nästa fråga om vi skulle anse det ändamålsenligt eller klarläggande att tillskriva datorer förmågan att göra bedömningar och utföra handlingar – kort sagt att ta ansvar. ”Det var datorns fel” kan vi ibland säga, men då uttrycker vi oss vanligen metaforiskt. Vi skulle inte dra det här uttryckssättet till sina yttersta konsekvenser: vi skulle t.ex. inte på allvar överväga att klandra eller straffa en dator. Vill vi dra någon inför domstol blir det snarast datorns tillverkare eller programmerare. Hur vore det att fullt ut leva med övertygelsen att datorer är subjekt?

Lars Hertzberg är professor emeritus i filosofi vid Åbo Akademi och medlem av Ikaros’ redaktionsråd.