OM INTELLIGENSEXPLOSIONENS VARA ELLER ICKE VARA

På senare tid har framstående forskare höjt varningsflaggor för att vi kan vara på väg att skapa ett självförbättrande system som utvecklas till en ”superintelligens”. En sådan superintelligens skulle då vara något vi inte kan kontrollera, och eventuellt kunde den utgöra ett hot mot mänskligheten. **Karim Jebari** och **Joakim Lundborg** undersöker förutsättningar för att en superintelligens skall uppstå och ser flera hinder i verkligheten.

Text: Karim Jebari och Joakim Lundborg

En uppfattning som fått alltmer uppmärksamhet är att generell artificiell intelligens (AI) skulle utgöra en av de största riskerna mot mänskligheten fortsatta existens. **Nick Boström**, en av världens främsta filosofer inom tillämpad etik, diskuterar dessa idéer ingående i boken *Superintelligence*, som ges ut på svenska i år. Två center för forskning om framförallt AI-relaterade risker har bildats vid universitetet i Cambridge, Centre for the Study of Existential Risk, (CSER) och i MIT, Future of Life Institute (FLI). Dessa har nyligen attraherat stora summor från privata och offentliga finansiärer, och i Sverige har professor **Olle Häggström** nyligen skrivit en bok om bland annat AI och annan teknologisk risk som kommer att ges ut på Oxford University Press. Forskning om AI-risk har med andra ord goda framtidsutsikter, trots att den ännu är i sin linda. Men är AI verkligen en av de stora riskerna? Det finns skäl, menar vi, att vara skeptisk.

Innan vi tittar närmare på Boströms argument är det viktigt att påpeka att ordet “intelligens” i det här sammanhanget inte är ett samlingsnamn på de funktioner som en hjärna kan utföra. I den här artikeln menar vi att “intelligens” bör förstås som förmågan att lösa problem oavsett om den förmågan förutsätter teknologiska hjälpmedel eller ej. Vidare är det här en förmåga som en person, en dator, ett företag eller ett samhälle kan ha. Det innebär att när vi pratar om “intelligens” i det här sammanhanget, så pratar vi om vad en entitet kan göra med all tillgänglig teknologi vid ett tillfälle, inklusive Internet och AI-system (till exempel Google).

Vad är då Boströms argument för att AI utgör en allvarlig risk för mänskligheten fortsatta existens? Sammanfattningsvis:

I. Människor är intelligenta, och vår intelligens är ett naturligt fenomen.

II. Mänsklig teknologi har lyckats återskapa, helt eller delvis, många naturliga fenomen.

III. I framtiden (om år, decennier, sekler) kommer mänsklig teknologi att skapa en artificiell intelligens (AI) med en intelligens jämförbar med en genomsnittlig människa (utan tekniska hjälpmedel).

IV. Denna AI kommer att på kort tid (timmar, dagar, veckor, månader) bli mer intelligent än någon existerande människa. Vi kan kalla detta för AI+. Eftersom “intelligens” = förmågan att lösa problem, och skapandet av en bättre AI+ är ett problem, så kommer AI+ att kunna skapa en AI++.

V. Steg (IV) kan upprepas ett antal gånger över mycket kort tid (minuter, timmar, dagar, veckor) tills det existerar en artificiell superintelligens (ASI), vars intelligens överstiger hela den mänskliga civilisationens kollektiva intelligens.

> Vår effektiva intelligens både som individer och som samhälle överstiger vida det som en mänsklig hjärna i sig är kapabel till på egen hand.

VI. När en ASI väl existerar kan den inte kontrolleras, och mänsklighetens öde kommer att ligga i dess händer, av samma skäl och på samma sätt som mindre intelligenta djurs öde ligger i våra händer.

Vårt huvudargument mot Boströms resonemang har följande form.

I. Påståendet att AI är en allvarlig existentiell risk förutsät-

ter en intelligensexplosion, alltså en mycket snabb övergång från AI –> AI+ –> AI++ –> ASI.

II. En intelligensexplosion är osannolik.

alltså:

III. Det är osannolikt att AI är en risk för mänsklighetens fortsatta existens.

Varför förutsätter resonemanget en intelligensexplosion?

Det är inte intelligens i absoluta termer som är farlig hos en ASI, utan intelligens i relativa termer, det vill säga det som är intressant är hur intelligent en ASI är *i förhållande till* de människor som en fientlig ASI kan vilja skada. Boström kallar detta "*decisive strategic advantage*". För enkelhetens skull kommer vi att med "AI", "AI+", "AI++" och "ASI" syfta på intelligenser i relation till vår intelligens, inte till den intelligens hos de människor som kommer att försöka kontrollera en sådan entitet.

Boström och andra AI-forskare tycks förutsätta att intelligensen hos de människor, företag och myndigheter som försöker kontrollera en AI inte kommer att förändras över tid. Det här antagandet är möjligen korrekt med avseende på vad vi normalt sett menar med "mänsklig intelligens". Men enligt den definition av "intelligens" som vi använder här, så kan en människa eller grupp människors intelligens höjas väldigt mycket av modern teknologi. En person kan till exempel lösa matematiska problem snabbare med en miniräknare, eller lagra större mängder information med en digitalkamera eller dator. Sofistikerade statistik- och sökprogram tillåter problemlösning som skulle vara omöjlig för personer utan tillgång till dessa verktyg. Vår effektiva intelligens både som individer och som samhälle överstiger vida det som en mänsklig hjärna i sig är kapabel till på egen hand.

Boström resonerar kring olika *takeoff*-hastigheter, det vill säga där det under olika scenarier tar olika lång tid att nå till en sådan acceleration att skillnaden blir tillräckligt stor. Enligt de modeller han ritar upp spelar det ingen roll hur lång tid det tar innan vi når *takeoff*, eftersom en exponentiell tillväxt, oavsett hur långsam den är, ändå vid någon tidpunkt kommer att korsa en kurva som inte växer exponentiellt. Vi menar dock att han inte tar hänsyn till att en accelererad AI-förbättring troligen påverkar intelligensen i det omgivande samhället. Av den här anledningen anser vi att det är nödvändigt med ett scenario som Boström kallar *fast takeoff* – en intelligensexplosion som är så hastig att dess landvinningar inte hinner spridas till resten av samhället innan ASI har uppnått en oöverkomlig fördel.

Om övergången från en AI via en AI+ till en ASI sker via en process som tar år eller decennier är det sannolikt att teknologiska framsteg som ökar människors och mänskliga organisationers intelligens hinner sprida sig genom samhället. Det innebär att människors eller mänskliga organisationers intelligens kan öka parallellt med utvecklingen av AI. När en AI+ (alltså en AI som är mer intelligent än någon människa är nu) väl skapas, så är det rimligt att den inte är mer intelligent än då existerande människor utrustade med den allra senaste teknologin, inklusive AI. I och med detta har en ASI inte längre en oöverstiglig fördel i jämförelse med dessa människor.

De som befarar en snabb utveckling av en ASI (en så kallad "intelligensexplosion"), tycks hysa uppfattningen att ett fåtal teoretiska insikter i AI-teori på kort tid skulle kunna resultera i stora framsteg. Men en sådan utveckling är sällsynt, inte bara i AI- utvecklingens historia, utan även i teknikutvecklingens historia. Till skillnad från populära narrativ om "stora språng", präglas teknikens och vetenskapens historia av misslyckanden, irrvägar och dyrköpta framsteg. Detta är knappast ett omöjlighetsargument, men det talar mot den här uppfattningen.

Boström och andra tycks anta att svårigheten att öka intelligensen i ett system ökar linjärt snarare än exponentiellt. En exponentiell ökning av svårigheten skulle innebära att även en ASI skulle ha stora svårigheter att förbättra sin kod. I modern datalogisk forskning har vissa problem redan lösts på ett sådant sätt att de inte kan lösas på ett (mycket) mer effektivt sätt. Ett exempel på ett sådant problem är "*edit distance*", ett problem där två sekvenser data analyseras i termer av hur den ena sekvensen kan transformeras till den andra. Nyligen presenterades ett bevis på att den lösningen som idag används är den bästa möjliga. Det innebär att oavsett hur intelligent en ASI än är, så kommer den inte att kunna lösa vissa typer av problem bättre än vad vi kan.

MAYBE ANOTHER QUOTE?

Ett annat skäl till att optimeringen av en AI kan vara mycket svårare än vad Boström och andra förutsätter är att de utgår från att en AI:s kognitiva struktur är transparent, alltså att det är relativt enkelt att se hur en AI löser ett problem genom att följa de operationer som den gör. Denna typ av transparens är dock bara aktuell i en viss typ av AI-forskning. I en stor del av de mest lovande samtida AI-systemen används en teknik som kallas "artificiella neuronnät". Dessa är kluster av simulerade neuroner som processar information på ett sätt som liknar en hjärnas. Neuronnät har framgångsrikt kunnat lösa en rad problem som mer transparenta

AI-algoritmer inte kunnat. Problemet med neuronnätverk är att de inte är transparenta. Det är alltså inte lätt för en programmerare (och sannolikt för ett självmedvetet neuronnätverk) att veta hur det fungerar. I frånvaron av transparens, är den formen av upprepad självförbättring som Bostrom befarar mycket svår. En AI+ som är ett neuronnätverk kommer att få mycket svårt att radikalt förbättra sin intelligens utan att först förstå hur dess kognitiva processer fungerar, ett mycket svårare problem.

Boström och andra forskare som oroar sig för en intelligensexplosion brukar också lyfta fram det faktum att processorer blivit allt kraftfullare och billigare. En AI skulle kunna dra nytta av denna trend för att med ren "brute force" (alltså med ren beräkningskapacitet) överkomma en del av den ökande komplexitetsbördan. Även detta argument är sårbart för de komplexitetsproblem som beskrevs ovan. Inte ens med en dator lika stor som en planet skulle en AI kunna lösa vissa typer av problem. Dessutom är det tveksamt om den intelligens som ren beräkningskapacitet ger är ett särskilt stort hot. Att kunna tänka snabbt är förstås en tillgång, men bara för en viss typ av problem.

En möjlig invändning mot vårt resonemang är att systemet människa + AI-system skulle vara mindre intelligent än endast ett AI-system, alltså att mänsklig intelligens skulle vara en negativ nettotillgång. Det här är möjligt, men är i så fall något som behöver motiveras. Boström nämner att mänskliga organisationer minskar i effektivitet eftersom systemet som helhet och människorna som ingår i det har olika agendor, till skillnad från en organisation bestående av enheter med identiska viljor. Vi anser att det här snarare rör sig om en *trade off* mellan flexibilitet och effektivitet. De olika agendorna gör det möjligt för organisationen som helhet att hantera många olika omständigheter, medan en organisation bestående av kopior skulle vara sårbar om det visade sig att den enade viljan hade missat en viss omständighet. En AI som behöver kunna anpassa sig till de föränderliga situationerna som råder i verkligheten skulle sannolikt behöva vara tvungen att ge upp denna typ av effektivitet.

Givetvis är det här ett sannolikhetsresonemang; *om* intelligensen hos en AI växer tillräckligt snabbt står vi ändå inför en ASI med ett avgörande strategiskt övertag. Vi anser dock att de invändningar vi reser här gör det mycket mindre sannolikt att så är fallet.

Innebär detta att AI eller ASI inte är ett hot eller något vi bör oroa oss för? Knappast. Boströms förtjänst är att han visar hur intelligens är ett oerhört kraftfullt verktyg, som kan påverka mänsklighetens öde. Istället bör slutsatsen bli att riskerna associerade med AI är mer komplexa och osäkra än att en ASI en dag kommer att förinta mänskligheten för att vi står i vägen för den. Om intelligens är makt och den makt som AI-teknologi för med sig bara tillfaller ett fåtal, kan det i sig vara en enorm risk för ett öppet och demokratiskt samhälle. Att en intelligensexplosion är osannolik innebär alltså inte att forskning om AI och teknikrelaterad risk är irrelevant.

Karim Jebari är doktor i filosofi och arbetar vi Institutet för framtidsstudier.

Joakim Lundborg är språkteknolog och arbetar på Wrapp.