Thore Husfeldt

Monstret i Turings bibliotek

Den svenske oxfordfilosofen Nick Bostrom presenterar i boken *Superintelligence* (2014) en dystopisk bild av stark artificiell intelligens: denna teknologi kommer inte bara att vara mänsklighetens största uppfinning, den kommer också att bli den sista, och måste anses som en omedelbar och katastrofal risk för vår art, lik molekylär nanoteknik, kärnkraft eller kemisk krigföring.

Artificiell intelligens med supermänskliga förmågor kommer att bli Homo sapiens sista uppfinning, eftersom alla efterföljande uppfinningar kommer att göras av ”superintelligensen” själv. Dessutom, om vi inte är ytterst försiktiga eller tursamma, kommer superintelligensen förgöra oss, eller åtminstone radikalt förändra våra livsvillkor på ett icke önskvärt sätt. Eftersom vi idag utvecklar teknologi som kan leda till en superintelligens, skulle just nu vara en bra tid för eftertanke.

Ingen vet, eller är överens om, hur man definierar intelligens, vare sig man pratar om allmän, artificiell, eller stark. Inte heller Bostrom. Han erkänner själv att hans bok till sin natur är spekulativ och förmodligen felaktig. Kanske är inte ens grunderna för den relevanta teknologin kända i dag. Men många av Bostrom argument är ganska robusta när det gäller detaljerna om ”vad?” och ”hur?”, och boken kan avnjutas utan en rigorös definition. Tills vidare kan vi föreställa oss superintelligensen som en hypotetisk agent som är mycket smartare än de bästa existerande mänskliga hjärnorna inom varje kognitiv aktivitet, bland annat vetenskaplig kreativitet, vanlig klokhet och sociala färdigheter.

VÄGAR OCH KONSEKVENSER

Bostrom beskriver ett antal vägar till superintelligensen. De består av aktuella forskningsområden och ny teknik, extrapolerade bortom människans kognitiva förmåga. Några exempel är

- neuroanatomiska metoder, såsom simulering av hela hjärnor (Einsteins hjärna i en näringslösning eller simulerad på en dator, neuron för neuron),

Filosofisk tidskrift 2015 nr 4, 44–53

- genetiskt eller artificiellt modifierade människor (hjärnimplantat, föräldrar som väljer könsceller för intelligens),
- intelligens som en emergent egenskap i enklare nät (internet som helhet blir ”medvetet”),
- datavetenskapliga metoder som maskininlärning (IBM:s Jeopardy-vinnande *Watson*) och ”gammal god artificiell intelligens” med symboliska resonemang (schackdatorer).

Ingen av dessa tekniker är idag i närheten av ens en dumboms intelligens, men på en historisk tidsskala är de alla väldigt nya. Bostroms beskrivning är en underhållande rundtur i datavetenskap, neuroanatomi, evolutionsbiologi, kognitiv psykologi och närliggande områden, informativ, men utan alltför mycket detaljer.

Konsekvenserna av att några av dessa forskningsprogram faktiskt uppnår sina mål är ännu mer spekulativa. Flera futurister har utvecklat scenarier för hur en superintelligent framtid kunde se ut, och Bostrom ger en överblick över många av dessa idéer. I en av dessa visioner överträffar accelererande och självförbättrande digitalteknik snabbt människans kognitiva förmågor och förvandlar vår miljö, precis som vi förvandlade den förhistoriska jorden. Snart är alla de protoner, som idag utgör planeterna i solsystemet, använda i bättre syfte som soldrivna datorenheter, kretsande runt solen i en stor nätverksintelligens kallad en dysonsfär. Det finns många andra scenarier, delvis beroende på vilken metod som kommer att visa sig vinna kampen mot vår egen intelligens, och om konsekvenserna kan tämjas. Inte alla dessa möjliga framtider är dystopiska. Vissa lämnar till och med utrymme för människor, kanske överförda till en digital version av odödlighet, eller som djur i djurparker. Men de innebär alla dramatiska förändringar i vår livsstil, jämförbara med införandet av jordbruk eller den industriella revolutionen.

KONTROLLPROBLEMET

Bostrom börjar sin bok med en kort fabel, där en grupp fåglar bestämmer sig för att leta efter en uggleunge som kan hjälpa dem med bobygge och andra mödor. Vi inser omedelbart deras dårskap: uggleungen kommer snabbt växa om fåglarna, kasta av sig deras ok, och förmodligen äta upp dem, i enlighet med sin natur. Det hade varit bättre om fåglarna hade tänkt över konsekvenserna av sitt sökande innan det var för sent.

I sin mest ofarliga form är superintelligensen dock bara en enhet som är väldigt bra på målinriktat beteende, utan egna önskemål, medvetande eller rovdjurstendenser. Men Bostrom visar hur även en sådan

superintelligent tjänare, väldigt foglig jämförd Hollywoods skenande mördarrobotar, skulle vara en fruktansvärd sak.

Bostrom beskriver ett intressant tankeexperiment om hur man anger beteendet hos en superintelligent maskin som producerar gem. Redan här finns det åtminstone två oönskade resultat som består i ”befängt förverkligande” av maskinens uppgift. Ett är att maskinen skulle kunna förvandla *allt*, inklusive människor, till gem. Eller, om man ger den en mer noggrant formulerad uppgift, att maskinen angriper problemet med att först öka sin egen intelligens (för att bli en bättre gem-producent), och gör om hela solsystemet till mikroprocessorer. Allt eftersom vi försöker göra vår order mer och mer precis, så fortsätter maskinen undvika att ”göra som vi vill” på alltmer intrikata sätt, med katastrofala resultat.

Jag tycker om att tänka på den här idén i termer av Goethes *Zauberlehrling*, förevigad av Musse Pigg i Disneys *Fantasia*. I denna berättelse beordrar trollkarlens lärling sin förtrollade kvast att fylla vatten i ett badkar. Lydig och bokstavstrogen drar den magiska tjänaren upp hink efter hink, men fortsätter tyvärr långt efter att badkaret har fyllts. Katastrofen kan avvärjas bara för att trollkarlen själv anländer i grevens tid för att rädda Musse från att drunkna. Ingen ondska var inblandad; i motsats till Bostroms allegoriska uggleunge har kvasten ingen annat mål än det pliktskyldiga utförandet av Musses order. Det var Musse som misslyckades med att vara tillräckligt precis. Hade Goethes kvast varit superintelligent, skulle den nog bara ha dödat lärlingen och kastat hans kropp i badkaret. Kroppen består ju till 70 procent av vatten, så uppgiften vore därmed slutförd både snabbt och elegant.

Man kan uppfatta sådana formalistiska lekar som banala övningar i att medvetet missförstå order. Men varje programmerare och lagstiftare vet att det är precis anledningen till att datorprogram är så svåra att skriva och lagar så svåra att formulera. Om det går att missförstå instruktionerna, så kommer de att missförstås.

Således finns det ingen anledning att anta någon illvilja hos superintelligensen; det räcker med en helt lydig agent som bara försöker utföra skenbart oskyldiga uppgifter. ”AI:n hatar dig inte, ej heller älskar den dig. Men du är gjord av atomer som den kan använda för något annat” (Yudkowski, 2006). Naturligtvis vore en superintelligens med egna önskemål, intentionalitet eller fri vilja ännu svårare att kontrollera.

Kontrollproblemet har många aspekter, varav en är rent institutionell: om superintelligens utvecklas i hård konkurrens mellan företag eller militärer, då är ingen av dem motiverade att stoppa utvecklingen, av rädsla för att förlora en teknologisk kapprustning. Som civilisation har vi tyvärr en föga imponerande meritlista för att förhindra globalt skadligt beteende när enskilda närliggande vinster finns inom räckhåll. Detta är

även relevant för den populära frågan ”Varför inte bara dra ur sladden?” För det första kommer det kanske inte finnas en sladd att dra ur.[1] För det andra, vem är ”vi”?

Men anta att vi kan lösa den institutionella frågan om vem som får utveckla superintelligensen. Då återstår att kodifiera våra mål och värderingar, så att den inte kommer att agera på ett sätt som vi finner icke önskvärt. Detta är naturligtvis ett problem som är lika gammalt som filosofin själv: Vilka *är* våra värderingar? Och återigen, vilka är ”vi”: arten, individen, eller våra efterkommande? Skall gemmaximatorn se till att inga barn dödas? Skall heller inga ofödda barn aborteras? Är det i alkoholistens intresse att få tag på alkohol eller inte?

Bostrom funderar över om vi kan specificera ”vad som är bra för oss” på en högre nivå, dvs. be superintelligensen själv att extrapolera våra önskemål baserade på en välvillig (snarare än bokstavlig) tolkning av våra egna bristfälliga beskrivningar, möjligen i en hierarki av beslut som skjutas upp till allt högre moraliska och intellektuella instanser. Förutsatt att vi skulle komma överens om meningsfulla ”robotikens tre lagar” som exakt beskriver våra värderingar utan att riskera befängt förverkligande, hur kan vi då motivera ett enormt överlägset intellekt att följa dessa lagar? Jag tycker att dessa spekulationer är både underhållande och tankeväckande.[2]

Vissa futurister välkomnar tanken om en paternalistisk superintelligens som avlastar vårt stenåldersmedvetande från att lösa våra egna etiska frågor. Men när vi delar vår miljö med en superintelligens, kommer vår framtid att bero på dess beslut, ungefär som gorillornas framtid beror på mänskliga beslut. Om vi för närvarande designar superintelligensen har vi bara en chans att säkra att den behandlar oss lika bra som vi behandlar gorillorna.

TURINGS BIBLIOTEK

Bakom Bostroms bok ligger ett tyst antagande om medvetandet kallad funktionalism. Om vi till exempel kunde simulera de neurokemiska processerna i en hjärna, då skulle denna simulering i sig vara ett medvetande med upplevelser, vilja och förmåga att agera. När vi väl förstår funktionerna i hjärnan är underlaget på vilket simuleringen utförs oväsentligt.

[1]Till skillnad från i filmvärlden måste vi anta att en superintelligens kan förutse de hinder som en grupp av morska hjältar kommer att lägga i dess väg. Bostrom beskriver olika sätt bli immun mot ett sådant angrepp genom att distribuera beräkningsunderlaget. En superintelligens kan säkert hitta på flera idéer.

[2]Jag gillar särskilt tanken att designa en AI med ett sug efter kryptografiska pussel där vi redan känner svaret.

Ur detta perspektiv blir frågor om kognition, intelligens, och agerande, ytterst till frågor om *beräkningar*.

Låt mig bjuda er till fiktiv plats som jag kallar Turings bibliotek. Den är inspirerad av *La Biblioteca de Babel*, som beskrivs i en novell av Jorge Luis Borges. Hans bibliotek består av en enorm samling av alla möjliga böcker i ett visst format. Nästan varje bok innehåller ren rappakalja, som om de hade skrivits av berusade apor. En bok består enbart av bokstaven A, medan en annan innehåller *Hamlet*.[3] I Borges berättelse befinner sig bibliotekarierna i ett konstant tillstånd av förtvivlan; omgivna av ett hav av kunskap, men oförmögna att navigera i det på grund av bibliotekets väldiga omfång. Våra hjärnor kan inte förstå dess storlek, och våra språk kan inte beskriva det,[4] ”ofattbart stort” kommer inte ens nära. Men biblioteksmetaforen hjälper oss att bygga någon form av mental bild av storleken.[5]

Föreställ er då Turings bibliotek. Varje volym innehåller programkod. Mycket av det är skräp, men vissa volymer innehåller meningsfulla program i något konkret programspråk, låt oss säga Lisp.[6] Vi kan mata in dessa program i en dator som finns i varje rum. De flesta skulle inte göra något nyttigt, många skulle få datorn att krascha. Från vårt startrum, fylld med bokhyllor, leder utgångar till andra rum i alla riktningar. Även dessa rum innehåller fler volymer, fler bokhyllor och fler utgångar. Det finns ingen klar ordning i biblioteket, men i närheten av oss ser vi en sliten volym som innehåller programmet

```
(print ”Hello, world!”)
```

som instruerar datorn att skriva det vänliga budskapet ”Hej världen!”, fast på engelska. Bredvid står en bok med den meningslösa frasen ”Etaoin shrdlu”, som inte är ett giltigt Lisp-program. Vi rycker på axlarna och ställer tillbaka den. Tack vare någon algoritmisk bibliotekarie hittar vi en tredje bok som innehåller den fullständiga koden för det enkla datorlingvistiska programmet *Eliza*, skrivet av Joseph Weizenbaum 1966. Detta berömda lilla program kan låtsas föra (skrivna) samtal, ungefär så här:

[3]En miljon andra böcker innehåller *Hamlet* med exakt ett tryckfel.

[4]Matematikens språk är ett undantag. Men våra hjärnor kan inte föreställa sig 10^{25}, antalet stjärnor, än mindre 10^{123}, antalet atomer i universum. Antalet böcker i biblioteket Babel är däremot $25^{1312000}$.

[5]Dennett (2005) har använt denna idé som ”Mendels bibliotek” för att illustrera den stora konfigurationsrymnden för genetisk variation.

[6]Jag valde Lisp på grund av dess roll som ett tidigt programspråk för AI, men andra programspråk skulle fungera lika bra. Snyggast vore det kanske att välja syntaxen i Alan Turings ursprungliga ”universalmaskin”. I beräkningsteoretiska termer är alla dessa modeller ”Turing-kompletta”: de kan simulera varandra.

– Berätta dina bekymmer.
– Min katt hatar mig.
– Varför tror du katt hatar dig?

Den primitiva grammatiska misstaget att utlämna ett pronomen är en ledtråd till att *Eliza* fungerar genom mycket enkla syntaktiska substitutioner, utan något försök att förstå vad du menar. Ändå är det bättre än att bara säga ”Hej världen!” Om vi fortsätter botanisera, kanske vi hittar koden för några system som var toppmoderna i början av 2010-talet, som IBM:s *Watson*, eller Apples *Siri*.

Idén är att varje algoritm finns någonstans i Turings bibliotek.[7] Om vi alltså accepterar funktionalismen, beskriver någon bok Einsteins hjärna: på frågor som ”Vad är 5 + 5?” eller ”Gillar du Brahms?” eller ”Bevisa Riemannhypotesen!” ger programmet samma svar som Einstein skulle ha gjort. Idag har vi ingen aning om hur ett sådant program skulle se ut. Kanske det bygger på symbolisk artificiell intelligens, likt *Eliza*. Kanske utgörs den största delan av ett ”konnektom” av Einsteins faktiska hjärna, dvs. en omfattande karta över dess neurala kopplingar. En sådan karta har varit känd för *C. elegans*, en genomskinlig rundmask med 302 neuroner, sedan mitten av 1980-talet. Visst är Einsteins hjärna betydligt större, men principen är inte absurd.

Någonstans i Turings bibliotek måste det finnas en beskrivning av en beräkningsprocess som är vida överlägsen alla mänskliga hjärnor. Annars vore ju Einsteins hjärna den smartaste algoritmiska intelligens som tillåts av logikens lagar, eller nära nog. Med det verkar osannolikt, för att inte säga självbelåtet, att evolutionen på planeten jorden redan lyckats konstruera den ultimata problemlösningsanordningen på bara några miljoner generationer.

Enligt Bostrom skulle upptäckten av denna beskrivning få katastrofala följder. I biblioteket finns ett monster. Vår utforskning av det bör fortsätta med försiktighet snarare än iver.

ATT LETA EFTER MONSTER

Om vi nu antar att beskrivningen av superintelligensen existerar någonstans i biblioteket, återstår det bara att hitta den. För icke-dataloger verkar den sista biten kanske relativt trivial. Även en andefattig uttömmande sökning skulle hitta volymen i fråga. Och även om detta tillvägagångssätt inte är en särskilt frestande uppgift för en människa, så är

[7]Vissa program kan vara så stora att de behöver spridas över flera böcker, kanske genom att hänvisa till andra volymer som subrutiner precis som modern programvara organiseras i programbibliotek.

ju datorer bra på att utföra meningslösa, väldefinierade, repetitiva rutinuppgifter med hög hastighet utan att klaga. Således kommer vi förr eller senare att stöta på monstret; den enda frågan är om söktiden mäts i år, generationer eller på geologiska tidsskalor.

Men detta är ett felslut. Det är ett felslut på grund av exponentialfunktionens helt orimliga tillväxt, och den uttömmande sökningens maktlöshet.

För att inse detta, kan vi återigen betrakta de 23 symboler som utgör det vänliga ”Hej världen!”-programmet. Vi kunde ha stött på dem, i princip, genom att undersöka alla sekvenser av tecken som utgör Lispprogram.

Hur mycket tid skulle det ta? Det är en enkel övning i kombinatorik, när vi väl har fastlagt några detaljer – hur många symboler vi behöver, hur snabb datorn är, osv. Men resultatet blir en besvikelse.[8] Om en modern dator hade börjat denna beräkning när universum skapades, skulle det ha kommit till någonstans runt (print ”Hello,”). Man kan vinna några extra bokstäver genom att kasta miljontals datorer efter problemet. Men inte ens då har universum tillräckligt med resurser för att genomföra en uttömmande sökning efter även mycket enkla program. Antingen tar tiden slut för att stjärnorna slocknar, eller det blir slut på protoner att bygga datorer med. Beräkning är en resurs, exponentiell tillväxt är enorm, och universum är ändligt.

Bostrom ägnar nästan ingen uppmärksamhet åt dessa problem. Till exempel skriver han i sin diskussion av hjärnsimulation:

> Framgång i simulering av en pytteliten hjärna, som den i *C. elegans*, skulle ge oss en bättre bild av vad som krävs för att simulera större hjärnor. Vid någon punkt i den teknologiska utvecklingsprocessen, när tekniker för automatisk simulering av små mängder hjärnvävnad blir tillgängliga, reduceras problemet till skalning.

Men för mig är skalning hela problemet.

Man kan invända att denna observation är ett banalt, kvantitativ argument som inte upphäver det övertygande faktum att i princip kommer en uttömmande sökning förr eller senare stöta på det fullfjädrade ”Hej världen”-programmet, och så småningom den monstruösa superintelligensen. Men det är precis min poäng: man kan samtidigt acceptera funktionalism *och* vara likgiltig gentemot utsikterna för AI. Spekulationer om den förestående upptäckten av monstret i Turings bibliotek måste grundas i beräkningstänkande, där tillväxttakten av beräkningsresurser är en kärnfråga.

[8] Lispprogram skrivs i ett alfabet av högst 128 tecken. En modern dator med 109 operationer per sekund, som har kört sedan Stora smällen för 14 miljarder år sedan, skulle inte ens ha hunnit titta på samtliga 12813 program av längd 13.

Kan det finnas ett annat sätt att upptäcka superintelligens än uttömmande sökning? Absolut. Naturen har ju själv redan upptäckt ett sådant monster: *homo sapiens* hjärna. Den har tagit sig dit genom hundratals miljoner år med hjälp av mutationer och urval i miljöer som gynnade kognition. Det finns alltså ett spår som naturen har följt genom Turings bibliotek för att komma fram till Einsteins hjärna.[9] Vi vet bara inte hur man känner igen, och inte heller hur man effektivt följer denna gradient. I entusiasmen kring artificiell intelligens på 1960-talet trodde vi kanske att vi befann oss på ett sådant spår. Men i så fall har vi övergett det. Den nuvarande trenden inom artificiell intelligens, bort från symboliska resonemang och mot statistiska metoder som maskininlärning, som inte syftar till att bygga kognitiva modeller, tycker jag är en osannolik väg.

Vår nuvarande kunskap om gränserna för beräkning är att det finns många till synes oskyldiga beräkningsproblem som är beräkningsmässigt svåra.[10] Efter mer än en generation av mycket seriös forskning om dessa problem känner ingen någon bättre metod än just uttömmande sökning för många av dem. Fastän de fantastiska funktionerna i din mobiltelefon lär berätta en annan historia, är den viktigaste slutsatsen av en generations forskning om algoritmer dyster: för de flesta beräkningsproblem vet vi inte vad vi ska göra.

Låt mig upprepa att jag inte försöker utesluta *existensen* av ett monster. Jag bara påpekar att även om ett monster existerar, behöver vi aldrig stöta på det. Turings bibliotek är alldeles för stort, ingen har märkt dess hyllor eller satt upp användbara riktmärken och kartor. Vår intuition tar fel när den extrapolerar att vi snart kommer att ha kartlagt hela biblioteket, bara för att vi har sett mer av det än våra farföräldrar någonsin gjorde. För ett halvt sekel sedan kanske det fanns hopp om snabba framsteg inom uttömmande sökning eller symbolisk AI. Men idag vet vi hur svårt det är.

FORSKNINGSINRIKTNINGAR

Om det algoritmiska perspektivet ovan är korrekt, så finns det inget att oroa sig för. Superintelligensen är en underhållande fiktion, inte mer värd att uppmärksamma än en förestående invasion av utomjordingar,[11]

[9]Om vi fortsätter denna idé, då vore en väg till superintelligens just att artificiellt utveckla intelligens i simuleringer av kompetitiva miljöer som gynnar kognitiva förmågor. Då har vi flyttat det tekniska problemet från simulering av hela hjärnor till simulering av hela adaptiva miljöer, inklusive hjärnorna i dem.

[10]Detta påstående är formaliserat i olika hypoteser i teorin om beräkningskomplexitet, såsom ”P är inte NP” och Exponentialtidshypotesen.

[11]Här är hela analogin: vi har utforskat månen och snart även Mars. Alltså måste vi oroa oss över den nära förestående kontakten med rymdvarelser.

kusliga demoner från havsbotten eller en oväntad förekomst av magi. Frågorna som tas upp är till sin natur ovetenskapliga, och vi kan lika gärna oroa oss över överbefolkning på Venus eller hur många änglar som kan dansa på ett knappnålshuvud.

Men jag kan ha fel. Jag satsar ju den mänskliga civilisationens framtid på ett antagande om beräkningssvårigheten hos ett problem jag inte ens kan definiera. Jag tror att jag har goda skäl till detta, och att dessa skäl är fastare förankrade i vår nuvarande förståelse av beräkning än Bostroms extrapoleringar. Men om jag har fel och Bostrom har rätt, så kommer mänskligheten snart att vara död, om vi inte satsar på hans forskningsagenda. I likhet med Pascals vad kan det kanske ändå vara klokt att ta Bostroms oro på allvar.

Problemdomänen som öppnas upp av *Superintelligence* kan leda till legitim forskning inom olika områden, oberoende av den underliggande hypotesens rimligheten.

Ett exempel är problemet med att kodifiera mänskliga värden. Bostroms farhågor förser dessa frågor med operativ betydelse, eftersom en formaliserad etik skulle kunde vara ramen för den incitamentsstruktur som behövs för att lösa kontrollproblemet. Men en ”algoritmisk moral-lära” vore ett värdefullt ämne hur som helst. Våra digitala samhällen styrs ju alltmer av algoritmiska processer. Dessa processer är långt ifrån superintelligenta, men de är mycket inflytelserika. Algoritmer väljer våra nyheter, fastställer våra försäkringspremier och filtrerar våra vänner och potentiella makar. Hur vi får dessa beslut att stämma överens med mänskliga värden är en giltig fråga för både etiker och algoritmiker, oavsett om algoritmerna är författade av Googles mycket verkliga dataloger eller av en mycket hypotetisk paternalistisk AI.

En annan fråga i samband medkontrollproblem är den pågående forskningen om semantik och verifiering av programspråk. Detta har redan varit ett väletablerat delområde för datavetenskap i många decennier. Några av problemställningarna går tillbaka till Gödels ofullständighetsteorem från 1930-talet, som visar att formella system inte kan resonera om beteendet hos formella system. Det har visat sig vara ett mycket svårt problem inom logiken att specificera och verifiera beteendet hos även mycket korta bitar programkod. Det kan mycket väl vara så att frågor hur man tämjer superintelligensen kan undersökas med hjälp av dessa verktyg. Andra vetenskapliga områden kan innehålla liknande anslutningspunkter. Incitamentsstrukturer är en fråga för spelteori och algoritmisk mekanismdesign, ett forskningsområde mellan datavetenskap och ekonomi. Kanske blir beräkningsaspekterna på artificiell intelligens tillräckligt väldefinierade för att leda till fruktbara frågor för forskare inom området beräkningskomplexitet. Det finns säkert fler exempel. Men

huruvida superintelligens kommer att bli ett giltigt ämne för några av dessa discipliner är mycket oklart för mig. Det kan bero på rent sociala frågor inom dessa forskningsgrupper.

Det kan också bero på tillgången till extern finansiering.

Denna sista punkt får mig att spekulera om en mycket verklig potentiell effekt av Bostroms bok. Det har att göra med synen på artificiell intelligens i allmänhetens ögon. Forskningsetik är en politisk fråga som avgörs av allmänheten. Den mest hoppfulla framtidsutsikten för forskning inom AI just nu är konstruktionen av en allmän artificiell intelligens. Tänk om allmänheten inte längre såg detta som ett mål, utan som ett hot? Om Bostroms dystopiska perspektiv leder till ökad allmän medvetenhet om superintelligens katastrofala effekter, då skulle AI inte länge anges under ”Mål” i en forskningsplan, utan under ”Etiska överväganden”, som kräver offentlig granskning i likhet med klimatförändringar, känsliga medicinska persondata, hejdlösa supervirus, djurförsök och nuklearkatastrofer. Och dessa, till stor del politiskt motiverade krav, skulle själva motivera forskning om kontrollproblemet.

Vi kanske behöver fundera över konsekvenserna av att stöta på ett monster i Turings bibliotek, oavsett hur trolig denna händelse kan vara.

LITTERATUR

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Dennett, D. C. 1995. *Darwin's Dangerous Idea: Evolutions and the Meanings of Life*. New York: Simon & Schuster.

Yudkowsky E. 2006. ”Artificial Intelligence as a Positive and Negative Factor in Global Risk”.